

371.25 R89 (2)

Keep Your Card in This Pocket

Books will be issued only on presentation of proper library cards.

Unless labeled otherwise, books may be retained for four weeks. Borrowers finding books marked, defaced or mutilated are expected to report same at library desk; otherwise the last borrower will be held responsible for all imperfections discovered.

The card holder is responsible for all books drawn on his card.

Penalty for over-due books 2c a day plus cost of notices.

Lost cards and change of residence must be reported promptly.



PUBLIC LIBRARY
Kansas City, Mo.

Keep Your Card in This Pocket

turn

211613'47
22G19'48 24

— SEP 6 '49 71 —

MEASUREMENT AND
ADJUSTMENT SERIES
EDITED BY LEWIS M. TERMAN

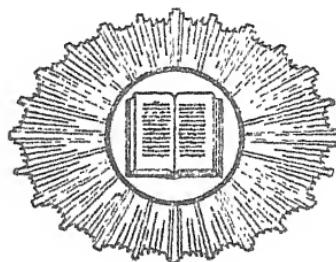
TESTS
AND MEASUREMENTS
IN HIGH SCHOOL
INSTRUCTION

BY G. M. RUCH

Professor of Education, University of California

AND GEORGE D. STODDARD

Assistant Professor of Psychology and
Education, State University of Iowa



Yonkers-on-Hudson, New York
WORLD BOOK COMPANY
2126 Prairie Avenue, Chicago

1927

WORLD BOOK COMPANY

THE HOUSE OF APPLIED KNOWLEDGE

Established 1905 by Caspar W. Hodgson

YONKERS-ON-HUDSON, NEW YORK

2126 PRAIRIE AVENUE, CHICAGO

Procrustes adjusted the stature of his guests to the bed which he provided, and it is not reported that any of his guests were so fortunate as just to fit the bed. Theseus ended Procrustes' career, and later, in Crete, found his way out of the Labyrinth by means of a fine thread. There is analogy in the fable of Procrustes to the procedure that formerly prevailed in our schools; but since 1917, when the Army tests were devised, testing has done much to banish Procrustean methods. Many tests have been prepared for the diagnosis and for the measurement of achievement, and a considerable body of writing on the subject of testing has come into existence. Indeed, so much material is now available that the teacher who would be well informed, but cannot specialize, has need for a guide. The present volume of Measurement and Adjustment Series is designed to supply guidance for instructors in high school subjects who would have none of the method of Procrustes. May it serve, like the thread of Theseus, to help them through the

Labyrinth

MAS : BSTMHSI-2

Copyright 1927 by World Book Company

Copyright in Great Britain

All rights reserved

PRINTED IN U.S.A.

PREFACE

THIS volume has been prepared primarily as a summary of the available tests and test methods in high school instruction. Since no separate treatment of measurement in secondary education has appeared to date, it was thought wise to include certain historical facts in Part I. In response to the increasingly critical attitude toward test selection, Chapter IV was prepared.

Part II presents the more important published measurements which are useful to the high school teacher. A great many tests have not been mentioned chiefly because of the lack of adequate data for their evaluation. Such tests are not necessarily undesirable, but no secure basis for their recommendation could be found in the literature. Part II presents numerous reliability coefficients, most of which were calculated by the authors or their students especially for this volume. Certain apologies might be offered for the scantiness of numbers of cases employed at times in the calculation of reliabilities, and for the many omissions in these data. However, it should be recognized that the burden of such determinations does not in fairness rest with the present authors, but rather with the authors of the tests themselves. If the reliabilities given are at times in serious error due to the limited numbers of cases used, or for other reasons, their presentation will at any rate stimulate their refutation and in time lead to the publication of more accurate results. Incidentally, it should be pointed out that very few makers of tests in the high school field have felt the necessity for ascertaining many pertinent statistical considerations which would be of the greatest value in the careful selection of tests.

In preparing Part II, the authors have tried to avoid being unduly swayed by such facts as great popular recogni-

tion of tests merely because of early historical origin, except to the extent that this is a legitimate criterion of merit. Where experimental findings justify the use of a less well-known test, such a test has been given as great recognition as others of more general reputation, or even greater.

Part III is to be justified upon the basis of the rapidly growing interest in informal objective examination methods. No treatment of test methods for classroom use can in the future ignore these unstandardized tests and examinations, which appear to be supplanting many of the traditional examination practices. Here again, especially in Chapter XVI, the authors were forced to take some position on a number of moot points. In the main, it is thought that the weight of the more extensive investigations indicates the fairness of the tentative recommendations advanced.

Part IV was placed last in the book, although logically it might have followed Part I. This was done for the sake of the reader who is not interested in statistical aspects of tests and test construction. The treatment of Part IV, moreover, is from the point of view of the test builder rather than that of the classroom teacher. A moment's thought will show that there is really no alternative. Part IV is to be defended upon two bases: (a) the fact that many high school teachers desire some insight into the more critical phases of test construction and selection; and (b) the fact that this book will doubtless find its way into the college classroom as a text in theoretical as well as applied aspects of measurement. There are so very few brief discussions of the general technique of test construction that it was felt that a few dozen pages could well be devoted to matters which obviously facilitate critical selection, use, and interpretation of measurements.

The authors' thanks are due a large number of graduate students who gave, scored, and tabulated test returns, and

who at times purchased test materials out of their own funds. Mr. Richard R. Foster calculated a considerable number of correlation coefficients for Part II. Grateful acknowledgment is made to Dean Paul C. Packer, of the College of Education, State University of Iowa, for funds used for the purchase and scoring of tests and for incidental clerical and stenographic assistance. Mrs. George D. Stoddard gave valuable aid in the preparation of certain sections of the book. Indirectly, the authors have drawn freely from the published work of many writers; particularly, from the statistical treatises of Dr. Truman Lee Kelley. Comment should be made upon the patience of the editor, Dr. Lewis M. Terman, and of the publishers in awaiting the completion of a task which has required the better part of four years.

G. M. RUCH
G. D. STODDARD

CONTENTS

	PAGE
EDITOR'S INTRODUCTION	xvii
PART ONE: STATUS, USES, LIMITATIONS, AND SELECTION OF TESTS IN SECONDARY SCHOOL INSTRUCTION	
CHAPTER	
I. THE PRESENT STATUS OF MEASUREMENT IN SECONDARY SCHOOLS	1
Historical — The first educational tests — The rise of intelligence tests — Other early educational tests — Educational testing at the present time — Limitations in the past: 1. Inertia; 2. College domination and teacher training; 3. Lack of sharp definition of content of high school subjects; 4. Qualitative <i>vs.</i> quantitative values in instruction	
II. USES AND LIMITATIONS OF TESTS IN THE HIGH SCHOOL	8
The major uses of standardized measures	
I. TESTS IN THE ADMINISTRATION AND SUPERVISION OF INSTRUCTION	8
Importance of tests in school supervision — Classification of supervisory uses of tests: 1. The measurement of pupil progress; 2. The diagnosis of teaching efficiency; 3. Setting up standards of performance; 4. The objectification of records of performance	
II. THE DIAGNOSIS OF SPECIAL DIFFICULTIES	18
The requirements for diagnostic tests	
III. TESTS IN GRADING, PROMOTIONS, AND SECTIONING	28
The general problems of classification — Tests in determining promotions — Sectioning of high school classes — The techniques of classification in the high school — Intelligence tests for classification — Educational tests for classification — Prognosis and aptitude tests — Educational tests <i>vs.</i> mental tests for classification	

CHAPTER	PAGE
IV. USES OF TESTS FOR RESEARCH PURPOSES	42
Tests in school investigations	
V. TESTS IN THE MOTIVATION OF LEARNING	43
Motivation as a neglected aspect of testing	
IV. CRITERIA FOR THE SELECTION OF EDUCATIONAL TESTS	45
General outline	
I. VALIDATION	48
Curricular validation	
II. RELIABILITY	51
Definition of reliability—The importance of reliability—What is a satisfactory degree of reliability?	
III. EASE OF ADMINISTRATION	56
Adequacy of instructions	
IV. OBJECTIVITY OF SCORING	58
Objectivity an essential in good tests	
V. INTERPRETATION OF RESULTS: NORMS	60
Accuracy of norms — Kinds of norms provided	
VI. DIAGNOSTIC VALUE OF THE TEST	64
The evaluation of the diagnostic functions of a test	
VII. NUMBER AND EQUIVALENCE OF DUPLICATE FORMS	65
The need for duplicate forms — Equivalence of forms	
VIII. TIME REQUIREMENTS, COST, MECHANICAL FEATURES, ETC.	66
PART TWO: DESCRIPTIONS OF HIGH SCHOOL TESTS BY SUBJECTS	
V. MATHEMATICS	71
Introduction	

Contents

ix

CHAPTER	PAGE
I. ALGEBRA	72
<i>Hotz First-Year Algebra Scales — Douglass Standard Diagnostic Tests for Elementary Algebra — Illinois Standardized Algebra Tests — Thurstone Vocational Guidance Tests: Algebra — Iowa Placement Examinations: Mathematics Aptitude, MA-1, Revised; Mathematics Training, MT-1, Revised</i>	
Remedial procedures in algebra	
II. GEOMETRY	84
<i>Minnick Geometry Tests — Schorling-Sanford Achievement Test in Plane Geometry — Hawkes-Wood Plane Geometry Examination — Geometry Test of Thurstone Vocational Guidance Tests</i>	
Remedial procedures in geometry	
III. OTHER MATHEMATICS TESTS	91
<i>Rogers Test of Mathematical Ability — Kelley Mathematical Values Test</i>	
Test materials — References	
VI. ENGLISH (LANGUAGE, GRAMMAR, SPELLING, READING, AND COMPOSITION)	97
I. LANGUAGE AND GRAMMAR TESTS	97
Introduction	
<i>Kirby Grammar Test — Wilson Language Error Test — Cross English Test — Pressey Diagnostic Tests in English Composition — Wakefield Diagnostic English Test — Tressler Minimum Essentials Test — Iowa Placement Examinations: English Aptitude, EA-1, Revised; English Training, ET-1, Revised</i>	
Remedial procedures in language and grammar	
II. SPELLING TESTS	108
Introduction	
<i>Sixteen Spelling Scales Standardized in Sentences for Secondary Schools (Seven S Spelling Scales) — Monroe's Timed Sentence Spelling Test, III</i>	
Remedial procedures in spelling	

Contents

CHAPTER	PAGE
III. READING TESTS	112
Introduction	
<i>Haggerty Reading Examination, Sigma 3</i> — <i>Thorndike-McCall Reading Scale</i> — <i>Van Wagenen Reading Scales</i> — <i>English Literature</i> — <i>Monroe Standardized Silent Reading Tests, III</i> — <i>Pressey Technical Vocabularys of the Public School Subjects</i> — <i>Holley Sentence Vocabulary Scale</i>	
Remedial procedures in reading	
IV. ENGLISH COMPOSITION SCALES	123
Introduction — What the composition scales measure	
<i>Hillegas Scale</i> — <i>Hudelson English Composition Scale</i> — <i>Hudelson Maximal Composition Ability Scales</i> — <i>Lewis Composition Scales</i> — <i>Van Wagenen English Composition Scales</i>	
Reliability of composition scales — Remedial procedures in English composition	
V. MISCELLANEOUS TESTS IN ENGLISH	130
Test materials — References	
VII. SCIENCE	136
Introduction	
I. GENERAL SCIENCE	137
<i>Ruch-Popenoe General Science Test</i> — <i>Dvorak General Science Scales</i> — <i>Van Wagenen Reading Scales: General Science</i>	
II. BIOLOGY	141
<i>Michigan Botany Test</i> — <i>Ruch-Cossmann Biology Test</i> — <i>Information Exercises in Biology (Cooprider)</i>	
Remedial procedures in general science and biology	
III. CHEMISTRY	146
Introduction	
<i>Powers General Chemistry Test</i> — <i>Iowa Placement Examinations: Chemistry Aptitude, CA-1, Revised</i> ; <i>Chemistry Training, CT-1, Revised</i>	
Other tests in chemistry	

Contents

xi

CHAPTER	PAGE
IV. PHYSICS	152
<i>Iowa Physics Tests — Hughes Physics Scales — Iowa Placement Examinations: Physics Aptitude, PA-1; Physics Training, PT-1 — Thurstone Vocational Guidance Tests: Physics</i>	
Remedial procedures in chemistry and physics	
Test materials — References	
VIII. FOREIGN LANGUAGE	160
I. FRENCH AND SPANISH	160
<i>Introduction</i>	
<i>Henmon French Test — Handschin Modern Language Tests — Columbia Research Bureau French Test — Iowa Placement Examinations: French Training, FT-1, Revised; Spanish Training, ST-1 — Wilkins Prognosis Test in Modern Languages — Iowa Placement Examinations: Foreign Language Aptitude, FA-1, Revised</i>	
Remedial procedures in French and Spanish	
II. LATIN	166
<i>Introduction</i>	
<i>Henmon Latin Tests — Ullman-Kirby Latin Comprehension Test — White Latin Test — Tests in Latin Vocabulary, Latin Derivatives, Latin Verb Forms, and Latin Syntax — Godsey Latin Composition Test — Orleans-Solomon Latin Prognosis Test</i>	
Comparative data on Latin tests	
Remedial procedures in Latin	
Test materials — References	
IX. SOCIAL STUDIES	177
<i>Introduction</i>	
<i>Barr Diagnostic Tests in American History — Pressey-Richards American History Test — Gregory Tests in American History, Test III — Van Wagenen Reading Scale in History — Kepner Background Tests in Social Sciences — Brown-Woody Civics Test</i>	
Test materials — References	

CHAPTER		PAGE
X. VOCATIONAL SUBJECTS	186	
I. MECHANICAL AND TRADE TESTS	186	
Introduction		
<i>Stenquist Mechanical Aptitude Tests — Thurstone Vocational Guidance Tests — Trade Tests</i>		
II. COMMERCIAL TESTS	191	
Introduction		
<i>Blackstone Stenographic Proficiency Tests, Typewriting</i>		
III. MUSIC TESTS	192	
<i>Seashore Measures of Musical Talent — Kwalwasser-Ruch Test of Musical Accomplishment</i>		
Remedial procedures in music		
Test materials — References		
XI. SURVEY TESTS	200	
Introduction		
<i>Iowa Comprehension Test — Iowa High School Content Examination — Iowa Placement Examinations — Columbia Research Bureau Tests</i>		
Remedial procedures in connection with survey examinations		
Test materials — References		
XII. GENERAL INTELLIGENCE TESTS	212	
Introduction		
Individual intelligence tests — Comparative data on group intelligence tests in the high school range — Reliability of intelligence tests — Intelligence test intercorrelations — Prediction of high school marks — Utilization of results of intelligence testing — Limitations of intelligence testing		
Test materials — References		
XIII. JUNIOR HIGH SCHOOL TESTS	227	
Introduction — The scope of measurement in the junior high school		
Description of tests for junior high school grades		

Contents

xiii

CHAPTER	PAGE
I. MATHEMATICS TESTS	228
<i>Buckingham Scale for Problems in Arithmetic, Division III</i> — <i>Compass Diagnostic Tests, Form A</i> — <i>Courtis Standard Research Tests, Series B</i> — <i>Monroe General Survey Scales in Arithmetic</i> — <i>Monroe Standardized Reasoning Tests in Arithmetic, Test III</i> — <i>Otis Arithmetic Reasoning Test</i> — <i>Spencer Diagnostic Tests in Arithmetic, III</i> — <i>Stanford Arithmetic Test</i> — <i>Stone Reasoning Test</i> — <i>Woody Arithmetic Scales</i> — <i>Woody-McCall Mixed Fundamentals, Forms I, II, III, and IV</i> — <i>Wisconsin Inventory Test in Arithmetic</i>	
II. ENGLISH TESTS	231
<i>Briggs English Form Test</i> — <i>Charters Diagnostic Language and Grammar Tests</i> — <i>Franseen Diagnostic Tests in Language</i> — <i>Ayres Spelling Scale</i> — <i>Iowa Dictation Exercise and Spelling Test</i> — <i>Stanford Dictation Exercise</i> — <i>Chapman-Cook Speed of Reading Test</i> — <i>Gray Standardized Oral Reading Check Tests</i> — <i>Monroe Standardized Silent Reading Tests, Revised</i> — <i>Stanford Reading Examination</i> — <i>Stone Narrative Reading Tests</i> — <i>Thorndike Test of Word Knowledge</i> — <i>Willing Scale for Measuring Written Composition</i>	
III. GEOGRAPHY TESTS	235
<i>Buckingham-Stevenson Place Geography Tests</i> — <i>Courtis Supervisory Geography Test</i> — <i>Gregory-Spencer Geography Tests</i> — <i>Hahn-Lackey Geography Scale</i> — <i>Posey-Van Wagenen Geography Scales</i>	
IV. VOCATIONAL SUBJECTS	238
<i>Ayres Scale for Measuring Handwriting, Gettysburg Edition</i> — <i>Courtis Standard Practice Test in Handwriting</i> — <i>Freeman Chart for Diagnosing Faults in Handwriting</i> — <i>Thorndike Handwriting Scale. For General Merit of Children's Handwriting</i> — <i>Gates-Strang Health Knowledge Test</i> — <i>Short Scales for Measuring Habits of Good Citizenship</i> — <i>Home Economics Information Tests</i> — <i>Murdoch Sewing Scale</i> — <i>Murdoch Analytic Sewing Scale for Separate Stitches</i> — <i>King-Clark Foods Test</i>	
V. SURVEY TESTS	239
<i>Illinois Examination</i> — <i>Lippincott-Chapman Classroom Products Survey Test</i> — <i>Otis Classification Test</i> — <i>Stanford Achievement Test</i>	

CHAPTER		PAGE
VI. INTELLIGENCE TESTS	...	241
<i>Illinois General Intelligence Scale—National Intelligence Test</i>		
Test materials—References		
PART THREE: INFORMAL OBJECTIVE EXAMINATION METHODS		
XIV. THE RÔLE OF INFORMAL OBJECTIVE EXAMINATIONS IN HIGH SCHOOL INSTRUCTION	...	251
Introduction—The relation between standardized and unstandardized tests—Limitations of the traditional examination—The subjectivity of the traditional examination—Unreliability of the traditional examination due to limited sampling		
XV. TYPES AND CHARACTERISTICS OF OBJECTIVE EXAMINATIONS	...	266
Classification		
Recall or Completion Tests—Multiple-Response Tests—The True-False Tests—Matching Exercises—Best Answer or Judgment Tests—Identification Exercises—Rearrangement Tests		
Advantages and limitations of objective tests in general		
Illustrations of the more common types of objective examinations		
XVI. CRITICAL CONSIDERATIONS IN OBJECTIVE EXAMINATION METHODS	...	282
Introduction—Corrections for chance or guessing—Instructions to guess or not to guess—Experimental data on guessing and corrections for chance—Conclusions from the preceding investigation—Results from similar studies—Time allowances for objective tests—Summary and conclusions		
Selected references for Part III		

PART FOUR: THE CONSTRUCTION OF EDUCATIONAL AND MENTAL TESTS

CHAPTER		PAGE
XVII.	THE CONSTRUCTION OF EDUCATIONAL AND MENTAL TESTS	301
	I. VALIDATION	301
	Introduction	
	Setting up a criterion of validity	
	The meaning of validity — The criterion or criteria of validity	
	Original selection of items — Methods: 1. Textbook Analysis; 2. Analysis of Courses of Study; 3. Analysis of Examination Questions; 4. Pooled Judgments; 5. Rating Scales; 6. Correlations with School Marks; 7. Rise in Percentage of Successes; 8. Correlations against Validated Measures; 9. The Method of Widely Spaced Groups; 10. Validation by the Principle of Social Utility; 11. Psychological and Logical Analysis; 12. Validation by Correlation with Other Measures	
XVIII.	THE CONSTRUCTION OF EDUCATIONAL AND MENTAL TESTS (Continued)	329
	The experimental try-out of items	
	Computation of item difficulties — Scaling or weighting of items	
	II. BREAKING THE TEST ITEMS INTO EQUIVALENT FORMS	338
	The second try-out — Determination of final time limits	
XIX.	THE CONSTRUCTION OF EDUCATIONAL AND MENTAL TESTS (Continued)	343
	III. THE DERIVATION OF NORMS	343
	The final try-out of the test — Types of norms: 1. Grade norms; 2. Age norms; 3. Percentile norms; 4. <i>T</i> -scores	

CHAPTER	PAGE
XX. THE CONSTRUCTION OF EDUCATIONAL AND MENTAL TESTS (<i>Continued</i>)	355
IV. DETERMINATION OF RELIABILITY OF THE TEST	355
1. The coefficient of reliability — 2. Measures of errors in individual scores	
V. PERFECTING THE ADMINISTRATION OF THE TEST	374
The Manual of Directions — The scoring keys	
Selected References for Part IV	
INDEX	377

EDITOR'S INTRODUCTION

IN the history of tests and measurements nothing stands out as more striking or significant than the recent increase of interest in this movement among teachers in secondary schools. The measurement of both intelligence and achievement began in the grades below the high school and for several years was rarely applied above the eighth grade. It is possible that the somewhat tardy introduction of measurement methods at the higher levels may have been due in part to a more marked tendency to conservatism on the part of high school teachers, who, at least in many parts of the country, have frequently had less pedagogical training than teachers in the elementary grades. Perhaps a more potent cause lies in the apparently greater difficulty of devising satisfactory measuring instruments for use with high school students. We are finding, however, that the supposed difficulties were more apparent than real. Experience shows that it is about as easy to measure either the natural abilities or the subject-matter achievement of high school pupils as it is to apply such measurements in the elementary grades. In some respects, indeed, it is easier. With older children a standard test procedure can be more rigidly followed and the factors of attention and effort can be more readily controlled. Moreover, there is probably no greater inherent difficulty in measuring achievement in algebra, geometry, physics, ancient history, or a foreign language than there is in measuring achievement in reading, arithmetic, or spelling.

Not only are measurement methods as applicable in the high school as elsewhere; in no part of the school system is more to be gained from their use. It is in the junior and senior high school that pupil guidance, both educational and vocational, assumes such outstanding importance. The student at this age stands at the threshold of life, subject to

conflicting desires and uncertain as to his abilities and as to the direction he should take. He is facing decisions which will affect his entire life: whether to leave school and go to work, whether to complete the high school course, whether to prepare for college, whether to fit himself to enter one or another of the professions. Without objective information of the kind which is obtainable only from standard tests, the guidance of such a student can rest upon little more than guesswork. One is tempted to put it more strongly and to say that *educational guidance without educational testing is professional quackery*, as much so as in the case of the physician who refuses to employ the approved laboratory techniques in the diagnosis and treatment of certain diseases.

That the truth of the above statements is coming to be generally recognized, among high school teachers as well as among university professors of secondary education, is indicated by the multiplication of high school achievement tests and by their widespread use. For most of the high school subjects fairly satisfactory tests have already been devised, and considerable progress has been made in the testing of special aptitudes and interests at this level.

One obstacle to the use of achievement tests in the high school has been the lack of a suitable textbook. The principal or teacher who would acquaint himself with the recent advances in this field has found it necessary to search through several educational magazines and innumerable monographs. In summarizing and interpreting the scattered contributions which have appeared to date in this field, the authors of the present volume have rendered a service to secondary education which the teaching fraternity will not be slow to appreciate.

The treatment of the subject throughout is direct, intelligible, and helpful. On the one hand it avoids over-emphasis of the theoretical problems of measurement, and on the

other hand it avoids too great preoccupation with the minutiae of procedure, scoring, and computations. The extensive experience of the authors (the senior author was formerly principal of the University High School at Eugene, Oregon) has enabled them to write a book ideally adapted to its purpose — one that is informative, balanced, and helpful. Both the uses and the limitations of tests are clearly and convincingly set forth. Criteria are given for the selection of tests suitable for a particular purpose. All the important intelligence and achievement tests intended for use in the high school are described and evaluated. Finally, in order that those who make use of testing methods may have more than a superficial knowledge of the instruments they employ, four chapters are added on the principles of test construction. In the opinion of the editor, no other chapters in the book are more worthy of careful study than these last four.

It is safe to predict that *Tests and Measurements in the High School* will find a wide field of usefulness both as a classroom text and for reading-circle purposes. It is a book which the progressive high school teacher cannot afford to ignore.

LEWIS M. TERMAN

PART ONE

**STATUS, USES, LIMITATIONS, AND SELECTION OF
TESTS IN SECONDARY SCHOOL INSTRUCTION**

TESTS AND MEASUREMENTS IN HIGH SCHOOL INSTRUCTION

CHAPTER ONE

THE PRESENT STATUS OF MEASUREMENT IN SECONDARY SCHOOLS

Historical. The introduction of educational and mental testing into the high school has not shown the same rapid growth which characterizes the applications of standardized measures in elementary education. A great many factors have been responsible for this situation. The retarding influences, for the most part, appear to be transitory in character, and present indications point toward an increasingly greater application and usefulness of test methods in secondary education in the future.

The technique of measurement, like that of curriculum construction, supervision of instruction, or the psychologizing of teaching methods, presents many unique problems in secondary schools which are not paralleled exactly in elementary school practice. Since the elementary school has in the past occupied the more strategic position in educational theory and administration, it is not surprising that the idea of the measurement of classroom products has had its most thorough try-out in the lower schools. Whether the elementary school will continue to lead in the test movement remains to be seen, but there is no doubt at present about the fact that test methods are gaining a secure foothold in the high school, college, and university. Viewed broadly over a period of years, the slower progress of measurement in the high school may prove to be a blessing in disguise, since it will enable the high school teacher or principal and the test builder to avoid many of the mistakes and pitfalls of the early years of experimentation with objective measures in the elementary

grades. Already there have emerged a number of ideas and principles which have been more or less adequately tested out and which can be adapted to the peculiar needs of the high school. The maker and the user of high school tests will also have the advantage of working with an educational public which has gradually been convinced of the general merit of standardized methods of measurement. That this is no small advantage is evident when we recall the storm of protest which greeted the statement of Dr. J. M. Rice thirty years ago, before the National Education Association, that the efficiency of the teaching of spelling could be measured by giving the pupils lists of selected words to be spelled.

The first educational tests. Rice is probably entitled to the credit of having produced the first educational test, for as early as 1894-95 he constructed two spelling "tests," one in list form and one in sentence form. Later he made similar tests in arithmetic and language. None of these was standardized in the present sense of the term, and no rigid attempt was made to insure validity or reliability.

The rise of intelligence tests. At almost exactly the same time (1894-95), Alfred Binet in France was at work on mental tests which proved to contain the germ of the first intelligence scale, the Binet scale of 1905. Binet's most important contribution was his 1908 scale, which introduced the concept of mental age.

The Binet scale was soon transplanted into America, but at first it had comparatively little influence on the rise of measurement in the schools, partly because it needed adaptation for American school children and partly because its earlier uses were too closely associated with the feeble-minded and institutional cases.

An American revision appeared in 1910, by Dr. H. H. Goddard at Vineland, New Jersey. However, the influence of the Binet tests in the public schools dates principally

from 1912-16 and the publication of the Stanford Revision of the Binet Scale by Dr. L. M. Terman. The publication of Terman's *The Measurement of Intelligence* in 1916 made the Binet scale easily accessible to the teacher and stimulated the training of Binet examiners throughout the country. Other American revisions of the Binet scale have been produced by Kuhlmann and by Yerkes, Bridges, and Hardwick, the last-mentioned authors using a somewhat different principle of scoring, known as the "point scale" method. The most recent revision is that of Herring.

The development of group tests of intelligence has expanded the use of intelligence tests enormously because of the economy introduced by testing large groups at one sitting in contrast with individual examinations necessitated by the Binet method. Just previous to the entry of the United States into the World War, Arthur S. Otis had been working at Stanford University under the direction of Terman on a test of intelligence which could be administered to large groups of persons at the same time. With the entry of the United States into the war, Otis's materials were placed at the disposal of the committee appointed to formulate mental tests suitable for the examination of soldiers. The Army Alpha tests were in considerable measure the result of adaptations of the Otis materials. The army tests served to popularize the intelligence-testing movement, and the close of the war saw a flood of group tests of intelligence — notably the Dearborn, Haggerty, Miller, National, Otis, Pressey, Terman, and Thorndike tests.

At the present time several million pupils are tested annually by group tests of intelligence, and the individual examinations by the Binet method run to hundreds of thousands. Similarly, the group test has resulted in industrial uses for the selection and placement of employees and in a wide variety of tests for civil service positions.

Other early educational tests. The first technical and scientific influence upon the rise of educational measurement is to be found in Thorndike's *Mental and Social Measurements*, which was published in 1904 and revised in 1913. This work provided the first presentation of the theory of educational measurement and placed the elements of statistical method within the grasp of the non-mathematically trained educator and psychologist.

During the years 1903-13, Thorndike and his students constructed a number of tests and scales which were validated and standardized; notably a handwriting scale (Thorndike, 1909), several arithmetic tests (Stone, 1908, and Courtis, 1909), a scale in English composition (Hillegas, 1912), a spelling scale (Buckingham, 1913), and later two reading tests (the Thorndike Visual Vocabulary Scales, 1914-16, and the Thorndike Scale Alpha 2 for Measuring the Understanding of Sentences, 1915-16).

Educational testing at the present time. Since 1915 the number of educational tests has increased by leaps and bounds. A study of several recent bibliographies leads to an estimate of nearly five hundred separate tests and scales, about one hundred and fifty of which apply to high school subjects. These numbers, however, have little meaning other than to point to the growing interest in quantitative methods in education, since probably considerably more than half of the tests are almost worthless and possibly not more than one fifth are deserving of use. Many are unstandardized, and comparatively few have been critically validated or examined for their reliability. A number of tests of wide popularity have later been shown to be of doubtful merit and utility. The test movement has suffered greatly from educational "faddism," with the result that many school administrators have been led to go through the movements of using tests in their schools under the mistaken idea that

wholesale testing was an indication of being scientific and modern.

In spite of these undesirable developments, there is no doubt that a saner and more critical attitude on the part of teachers and administrators is gradually arising; and with the introduction of courses covering testing programs, test construction, and statistical methods into the colleges and universities, the future is sure to witness the production of many tests and scales of great merit.

Limitations in the past. The slow growth of educational testing of high school subjects has been due to a variety of factors which have operated in the past, and some of which will continue to operate in the future. The more important of these factors will be presented in brief discussion.

1. *Inertia.* Because of the fact that educational theory of the past few years has held that the elementary grades are the critical ones in the life of the pupil, relatively less attention has been given to the problems of the high school. This statement applies to other aspects of education as well, especially to the development of the course of study, methods of administration and supervision, and the application of psychology to the methods of instruction. It must be admitted that justification for the focusing of attention upon elementary instruction has not been entirely lacking, but, with the rapid growth in the enrollment in secondary schools, the twelve-year course is coming to be the recognized public education for the average child. The next decade or two promises to be the period of scientific study of high school education.

2. *College domination and teacher training.* The high schools have been far more completely under the domination of the colleges and universities than have been the lower schools, with the result that the strictly professional training of high school teachers has been neglected in favor of academic

training in the subject matter to be taught. The high school teacher has been more of the specialist and more prone to assume the attitude that it is his business to know *what* to teach rather than *how* to teach. For these reasons supervision of instruction has not been stressed as a major function of the principal or superintendent. Since the use of tests is almost a parallel development to the supervision of instruction, it is quite natural that the high school has not been affected by either of these practices to the same extent as the elementary school.

3. *Lack of sharp definition of content of high school subjects.* With the possible exceptions of mathematics and Latin, there has been little agreement among educators either as to the objectives or as to the content of the high school subjects. This lack of uniformity has operated to make the production of tests of general usefulness very difficult for high school subjects. As a consequence, comparatively few tests have been developed in such subjects as history, modern languages, science, English, and the vocational studies. Such *content* subjects, in contrast with the *tool* subjects like arithmetic, reading, writing, and spelling, are admittedly more refractory to the efforts of the test maker. The result has been few tests, and these poorly constructed and inadequately validated.

4. *Qualitative vs. quantitative values in instruction.* High school teachers have been disposed to hold to a doctrine of values which presents many points of fundamental antagonism to the aims of educational measurement. The introduction of test methods in the measurement of the results of teaching has often been regarded as an impossibility because of the inability of tests to reveal appreciative, cultural, spiritual, and disciplinary values. Several logical fallacies may be pointed out in the position taken by such objectors.

In the first place, it is illogical to argue from what has been

done in the past to a conclusion that future developments are impossible. The building of adequate tests and scales always waits upon the clarifying of aims and objectives. It is both more hopeful and more reasonable to assume that the skill of the test maker will prove equal to that of the builder of curricula.

An even more serious error in logic resides in the attempt to divorce qualitative and quantitative aspects of knowledge. The subject of English, for example, does present certain values not to be found in a tool subject like writing, but to call the one qualitative and the other quantitative introduces far more loose thinking than it avoids. It has been stated, and with no little truth, that "whatever exists, exists in some quantity." It follows as a corollary that it can be measured.

The appreciation of literature, correct habits of citizenship, love of music and the other arts, and respect for public law and morality are unique products which education hopes to produce as its major objectives, but intangibility is not to be thought of as negating the possibility of their quantitative expression. And, as in the case of the question of mental discipline, the test critic and the test advocate alike, along with educators in general, should abide by the present tacit agreement to waive such questions of qualitative *vs.* quantitative values in education until such differences have actually been proved.

It is probably true that any field of knowledge which can be analyzed into its constituent unit skills, knowledges, and abilities to a degree that permits the teaching of its subject matter by a method better than rule-of-thumb can be measured by tests with validity as great as that underlying such analysis and such teaching methods.

CHAPTER TWO

USES AND LIMITATIONS OF TESTS IN THE HIGH SCHOOL

The major uses of standardized measures. Educational and mental tests have proved useful in at least five principal directions in high school teaching and administration. These applications may be classified for convenience as follows:

- I. Supervision and administration of instruction
- II. Diagnosis of special difficulties
- III. Grading, promotions, and sectioning of classes
- IV. Research purposes
- V. Motivation of learning

I. TESTS IN THE ADMINISTRATION AND SUPERVISION OF INSTRUCTION

Importance of tests in school supervision. The use of tests for the supervision of instruction has been the most important function of such measures and includes a wide variety of uses; viz., general surveys of teaching efficiency, studies of pupil progress, comparisons between school systems, comparisons of schools or classes within the same system, comparisons of the accomplishment of individual pupils with norms of average performance, etc. The more detailed uses of educational and mental tests in supervision shade over into the II and III divisions listed above.

City school surveys like those of Butte, Cleveland, Gary, and Salt Lake City have made extensive use of educational tests, the tests employed in some cases being constructed for the occasion.

The central feature of testing in supervision is the *reference to a norm or standard of pupil performance.* With minor

exceptions, all practices falling under this heading are, essentially, comparisons of pupils, classes, or schools with average or typical conditions. Such norms are most often mean (average) or median performances of relatively unselected pupils of given ages, grades, or school subjects. Mental tests ordinarily are provided only with age norms. Educational tests most often are provided only with grade norms. Theoretically, and practically, age norms are to be preferred over grade norms in educational tests also, since age norms are less likely to conceal conditions of faulty grade classification — i.e., over-ageness of pupils — and because they permit more direct comparisons with mental test scores. Both age and grade norms are limited in their usefulness in the high school because most high school subjects are not continuous over a period of years. Such subjects as English, grammar, reading, and foreign languages lend themselves to the use of grade norms, however. In general, subject norms are the most valuable in Grades IX to XII. For subject norms the use of percentiles and *T*-scores is the prevailing practice. (See Chapter XIX for the details of such norms.)

Reference to a norm or standard may be thought of as similar to the use of a "control group" in scientific experimentation. The pupil, class, or school system is compared in performance with a larger outside group which represents a balancing of the factors arising from differences in types of teaching, textbooks, geographic and local school conditions, teacher skill, pupil motivation, etc. For this reason, norms, at best, are very crude guides to what is a reasonable performance for a given pupil, class, city, or other unit.

Classification of supervisory uses of tests. The problems of supervision which offer opportunity for the application of mental and educational tests include the following:

1. The measurement of pupil progress
2. The diagnosis of teaching efficiency
3. Setting up of standards of performance
4. The objectification of records of performance

1. *The measurement of pupil progress.* The first of these supervisory uses of tests is exemplified by the practice of September and June testings. Due to the transferring in and out of a particular school system and to the normal loss of efficiency over the summer vacation, September-June comparisons may have very great value. By reference to the norms the supervisor is often enabled to discover whether the progress throughout the year is normal, below normal, or above normal. The value of test scores for pupils given conditional or special promotions is evident. September-June comparisons become more valuable if provision is made for individual record cards on which the scores are entered, preferably in graphic form. Such records should be made accessible to teachers, supervisors, and parents. A very desirable practice in connection with the keeping of educational test records would be the making of such records available to the pupil himself. This point, however, will receive attention in a later section of this chapter.

The objectivity of pupil record cards makes them particularly convincing evidence to the parents in cases of non-promotion, extra-promotion, assignment to special classes, etc. The cards should be designed to show both the norm and the actual performance of the pupil in a graphic manner.

The comparison of pupil, class, or school accomplishment with the norms of the test are interesting, often valuable, and with care can be made very meaningful. These comparisons are never simple, and a number of additional facts must be weighed simultaneously with the test scores if the proper interpretations are to be made. Such facts are:

- (a) The average age of the pupils tested
- (b) The general mental ability of the group tested
- (c) The previous preparation of the pupils (including time allotments, rural *vs.* city training, etc.)
- (d) The validity of the tests employed when compared with the aims and content of the curriculum of the particular school

The first two of these facts are capable of rather exact and objective statement. The last two are less tangible but nevertheless open to evaluation.

Neglect of such facts as age and mental age in the interpretation of test scores is probably the most widespread misuse of educational test records today. Teachers and superintendents very often go no farther in their interpretations of the score of a pupil or a class than to compare the median or mean performance with the norm. If the pupil or the class is above the published figure, the conclusion is drawn that satisfactory progress has been demonstrated. If the pupil or the class fails to reach the norm, it is concluded that the achievement is unsatisfactory.

The following set of data will provide a more objective basis for considering an important source of error in interpreting educational test results:

Average score of Grade IX pupils in School A	40.1
Average score of Grade IX pupils in School B	42.8
Norm for Grade IX	40
Average age of pupils used in norms	15-6
Norm for Grade X	45
Average age of pupils used in norms	16-6
Average age of pupils in Grade IX in School A	15-7
Average age of pupils in Grade IX in School B	16-5

At first glance, School B would seem to be rather definitely superior to School A, the difference being roughly half a

grade. However, when we examine the further data concerning the ages of the two ninth-grade classes and the average ages of the pupils entering into the norms, it is evident that School B's claim to superior accomplishment reduces to the situation: Given the premium of one year of additional age and maturity, School B does a quality of work about one half a grade better than School A. This situation reverses the claims of School B. The correct interpretation, in the light of all the facts given above, is that School A is almost exactly normal in attainment and that School B is markedly below normal (the normal attainment for pupils of an average age of approximately $16\frac{1}{2}$ years being tenth-grade, not ninth-grade, performance). Here, as often, *apparently superior accomplishment is in reality purchased by the very simple expedient of an abnormally large percentage of over-ageness of the pupils*. Of course, no school would deliberately fail to promote pupils in order to produce such results. Nevertheless, many schools give little or no close scrutiny to age-grade distributions, and it is not very unusual to find average ages of pupils in the same grade of different schools differing by six months to a year or more.

The situation is similar with respect to the influence of differences in mentality of pupils, classes, or schools when compared with one another or with general norms. Mental test scores, although far from perfect indices of learning capacities, do show by the correlations of mental ages with success in various school subjects the possibilities for error in comparisons of groups differing in mentality. For this reason alone it may be urged that *mental tests should ordinarily be given along with educational tests when refined comparisons are contemplated*. This can be shown by the illustration used for Schools A and B above, if we substitute "average mental age" for "average age" in the data.

The writer had occasion to study the results of a number

of educational tests and Binet tests of all the eighth-grade classes of a California city. Enormous differences were found in the medians of the educational tests, but the Binet IQ's showed corresponding differences in practically every case, the range of *median* IQ's of these eighth-grade classes being from 80 to nearly 110. In terms of mental maturity, the pupils of the best eighth-grade class were roughly four years older mentally than those in the poorest eighth-grade class. Knowing these facts, no particular alarm need be felt that very unequal school accomplishments were found among these classes. The example selected can be duplicated in almost any large city.

Inequalities in the kind and quality of previous instruction often complicate the task of interpreting test results. It is well known that rural children often fail to reach the same standard of efficiency on school tests as do city children. The reasons for this are not entirely clear, and the evidence upon which the statement is based is not as conclusive as might be wished. Possible explanations may be found in the difficulties attending the one-teacher school in contrast with the highly specialized and departmentalized teaching of modern city school systems. Secondly, it is possible that cities tend to attract the more intelligent elements of the population, and there may be a tendency for the brighter and more ambitious country children to attend city schools. Again, the school term is frequently shorter for the rural schools. Principals of high schools often report that pupils from certain elementary schools within the same school system uniformly make lower scores on the tests given for sectioning purposes than those from other elementary schools within the same system.

Tests themselves are not without serious limitations when applied to the measurement of pupil progress. Since any educational test must necessarily represent a very small

sampling of the total skills and abilities covered in even a single school subject, they are certain to be of limited validity in any particular school situation. The following five statements are roughly true and will serve to call attention to inherent limitations of all tests:

- (1) Educational tests are almost invariably *small samples* of the total function as actually taught.
- (2) They represent *general teaching practices* rather than the particular methods of any teacher or school system.
- (3) They cover only the *major aspects* of the subject and omit the less important objectives.
- (4) They are *not adaptable* to local conditions.
- (5) Their norms too often place their faith in *large numbers* rather than in careful observance of the principles of sampling; i.e., the norms are often based upon returns voluntarily mailed to the authors by users of the test rather than on pupils carefully selected so as to be representative of the various ages or grades.

For these reasons it is to be expected that educational tests at times minimize and at times exaggerate the gains produced by careful teaching. An example is found in reading, English, and language tests which seldom cover more than two or three aspects of the subject to the neglect of perhaps a dozen objectives in the teaching of such subjects. The only way out of these difficulties at present is the application of common sense to the problems of test interpretation, bearing in mind the actual scope and validity of the tests employed.

2. *The diagnosis of teaching efficiency.* This problem has many points in common with the discussion of the measurement of pupil progress. It is probably true that, all other things being equal, it is a valid procedure to measure

the efficiency of the teacher by the progress of her pupils. Actually, all other things are seldom equal. If two teachers are fairly to be compared by means of educational test scores of their pupils, it is necessary to establish a number of facts about the two classes. Some of the questions to be asked follow:

- (1) Has the previous preparation of the pupils in the two classes been substantially the same? This is practically equivalent to establishing the fact that the two classes show equal performances on valid tests at the beginning of the school year.
- (2) Do the two classes represent about the same level and range of talent with respect to mental ability (or learning capacity)? The best solution of this situation at present is the application of intelligence tests.
- (3) Are the classes comparable in age? Age-grade distributions of the classes to be compared will provide the needed data here.
- (4) Do the tests employed approximate the actual content and emphasis of teaching equally well for the two classes to be compared?

These questions do not exhaust the possibilities, but they are suggestive of the line of reasoning which must be taken. Used with caution, the good educational test does provide an objective and impartial estimate of teaching success which is greatly superior to the impressionistic method employed by supervisors in "sensing" good and bad teaching. It is certainly safer than the occasional "dropping in" for five minutes by the superintendent. Moreover, experimentation is now showing that educational tests are more reliable than the traditional written examinations which are the almost universal practice at the present time. The combined use of educational tests and informal objective

examinations (the so-called "new type" examinations), together with all the ascertainable facts about mentality, previous training, over-ageness, etc., will afford a practical solution of the problem of teacher evaluation.

3. *Setting up of standards of performance.* The supervisory functions in which test methods have a part include the determination of standards of reasonable performance for the pupil. Test norms have some value in this connection, but they represent, primarily, what average pupils *actually* accomplish rather than what pupils *might* accomplish under conditions of high motivation and superior teaching. Some test workers have advocated the establishment of several separate types of norms, as follows:

- (1) "Minimum accomplishment norms"; i.e., a critical point below which attainment should be considered of less than passing quality.
- (2) "Balanced norms"; i.e., norms of average or typical performance. These are the type supplied with educational tests at present.
- (3) "Motivated" norms; i.e., norms derived on pupils working up to the limits of their capacities.

The first type of norm is a practical impossibility, and subject to all the difficulties attending the setting up of criteria of minimum passing quality of school work.

The second type is the prevailing one at present and is open to the many objections already discussed.

The third type of norm might be derived at great expense and much experimental labor, but final proof of its validity would always be lacking. There is of course a physiological limit to improvement in any school subject, but the difficulties of adequate motivation and lack of knowledge of the best possible methods of teaching will always prevent a close approximation to such limits.

A partial solution of the problem of reasonable accomplishment may be found in the recently advocated *accomplishment quotient* or *achievement quotient*. This assumes the mental age of the pupil to be a more or less valid index to his learning capacity. The achievement quotient (AQ) is the following ratio:

$$AQ = \frac{\text{Educational Age}}{\text{Mental Age}}, \text{ or, more simply, } AQ = \frac{EA}{MA}.$$

There are a number of statistical objections to such ratios, but the major weakness seems to lie in the fact that it assumes the mental age to be a valid index of learning capacity. This is perhaps roughly true for such subjects as reading and mathematics, but it is far from true for such subjects as spelling, cooking, manual training, etc. A second practical objection arises, in the case of high school subjects, from the fact that age norms are not very accurate or serviceable in the upper grades. These objections also apply to the alternative AQ formula which uses the chronological age of the pupil as the denominator term, thus:

$$AQ = \frac{EA}{CA},$$

except that there is a certain defense for an EA/CA comparison as a matter of naturalness of thinking without implying any close connection between age and learning capacity.

The most important thing in all this discussion of reasonable standards of attainment is the recognition by test users of the principle that *no one norm of performance can be set up which will have universal validity for all pupils or all schools*. Psychological studies, no matter what directions they have taken, have shown enormous individual differences in human capacities. Single school grades commonly have from four to six grades, or years, of educational ability

represented within a single grade. This condition argues for two things: (1) closer and more homogeneous classification of pupils, and (2) interpretations of test results in the light of the known facts about the range of individual differences among school children. Here, again, common sense together with the thoughtful use of appropriate educational and mental tests must suffice for the present in determining reasonable standards of pupil accomplishment.

4. *The objectification of records of performance.* The educational advantages of impartial and objective records of accomplishment have already been referred to in relation to supervision. Two further advantages in such records center about the school-parent relation and the motivating effects upon the pupil. The latter will be discussed in a later section.

Standard test results place the teacher or principal in a more secure position in dealing with parents who are disinclined to admit the facts, as they often are when their children face situations of failure or have to be assigned to special classes. The explanation of the pupil's poor progress in the light of a carefully tabulated test record is not nearly so formidable a task as when the principal must rest his case on mere statements of the teacher to the effect that the pupil in question is doing "unsatisfactory" work. The parent can often be shown the impartiality of the test and the significance of its norm when prejudice against a teacher prevents unbiased consideration of the case apart from the personalities involved.

II. THE DIAGNOSIS OF SPECIAL DIFFICULTIES

The requirements for diagnostic tests. The term *diagnostic tests* has been used in a very loose and often misleading manner by test workers. Diagnosis is, of course, a relative term, but it is nevertheless capable of definition. Study of the many educational tests laying claim to diagnosis reveals

very few that can fully substantiate their claims. The future is sure to witness rapid developments in the production of genuinely diagnostic tests.

The following statements characterize genuine diagnosis as a practical test situation :

- (1) The school subject to be diagnosed must be broken into all of its important constituent unit skills or aspects, and each of these must be measured separately.
- (2) Each of these units must be sampled widely enough (i.e., be covered by enough test items) that no important facts or skills are omitted.
- (3) Each of these units (whether designated as "Parts" of the total test, or as separate tests) must be provided with separate norms for the interpretation of scores by units.
- (4) The score yielded by each unit of the total test must be reliable enough to stand alone as a score on an *individual* pupil in contrast with group measurement.
- (5) No tabulation of individual errors should be required in order to arrive at a diagnosis.
- (6) The analysis into units should be carried far enough that each unit parallels a unit in the course of study; i.e., represents some one teaching unit.
- (7) The diagnosis should suggest the remedial or corrective program which should follow the diagnostic testing.

It must be admitted that the foregoing requirements are very stringent, and it is doubtful whether any high school test published to date meets all of them very well.¹ Table 1 lists a number of high school tests or series of tests which

¹ One of the best examples of a series of tests which approximate these requirements is the *Compass Diagnostic Tests in Arithmetic* (Scott, Foresman & Co., 1925). About ninety so-called "unit skills" are isolated within the field of arithmetic in a series of twenty different tests.

TABLE 1
DESCRIPTIONS OF CERTAIN HIGH SCHOOL TESTS OF DIAGNOSTIC VALUE

NAME OF TEST	UNIT ABILITIES COVERED	TIME LIMITS	RELIABILITY COEFFICIENTS	P E _{98.1} S.D.	N
Barr Diagnostic Test in American History	1. Comprehension 2. Chronological judgment 3. Historical evidence 4. Evaluation of facts 5. Causal relationships Total	6' 6' 6' 6' 6' 30'	.63 .53 .30 .24 .62 .77	.32 .34 .31 .29 .33 .28	50
Douglass Standard Diagnostic Tests for First-Year Algebra	1. Addition and Subtraction 2. Multiplication 3. Division 4. Simple Equations Total, Series A	7' 8' 10' 9' 34'			
	5. Fractions 6. Factoring 7. Formulae 8. Simultaneous Equations 9. Graphs 10. Roots, Exponents, Radicals 11. Quadratic Equations Total, Series B	12' 15' 15' 15' 15' 15' 15'	.80 .63 (a)	.27 .45	175 1040

(a) As given by Douglass, *University of Oregon Publication*, Vol. II, No. 5 (June, 1924), page 19.

TABLE 1 (Continued)

NAME OF TEST	UNIT ABILITIES COVERED	TIME LIMITS	RELIABILITY COEFFICIENTS	$\frac{P.E. \cdot 0.1}{S.D.}$	N
Godsey Diagnostic Latin Comprehension Test	1. Translation 2. Rules of grammar Total	30'			
Handschin Modern Language Tests (French and Spanish)	1. Silent Reading 2. Comprehension and Grammar: French	5' 10'	.89	.21	86
Henmon French Tests	1. Vocabulary 2. Sentences Total	10' 10'	.61	.33	60
Henmon Latin Tests	1. Vocabulary 2. Sentences	10' 10'	.65-.80 .52-.71	.27-.92 .31-.34	44 44
Hotz Algebra Series B	Scales, 1. Addition and Subtraction 2. Multiplication and Division 3. Equations and Formulas 4. Problems 5. Graphs Total	40' 40' 40' 40' 40' 200'			
				.02	.18
					175

TABLE 1 (Continued)

NAME OF TEST	UNIT ABILITIES COVERED	TIME LIMITS	RELIABILITY COEFFICIENTS	P.E. ∞ :1 S.D.	N
Iowa High School Content Examination, Form A	1. English and Literature 2. Mathematics 3. Science 4. History Total	20' 20' 20' 20' 80'	.93 .93 .83 .89 .95	.17 .17 .26 .21 .15	247 247 247 247 247
Iowa Physics Tests	1. Mechanics 2. Heat 3. Electricity and Magnetism	45' 45' 40'			
Kirby Grammar Test	1. Grammatical choices 2. Principles of Grammar Total	35'	.55 .78	.38 .28	80 80
Pressey (and others) Diagnostic Tests in English Composition	1. Capitalization 2. Punctuation 3. Grammar 4. Sentence Structure	None (b) 5' 10' 15' 15'	.79 .78 .90 .73	.28 .28 .20 .29	99 99 99 99

(b) Estimated by Pressey as time for all but two or three slowest pupils of class to finish.

TABLE 1 (Continued)

NAME OF TEST	UNITS ABILITIES COVERED	TIME LIMITS	RELIABILITY COEFFICIENTS	P.E. _{so.1} S.D.	N
Pressey-Richards Tests in the Understanding of American History	1. Character Judgment 2. Historical Vocabulary 3. Sequence of Events 4. Cause and Effect Relations Total	5' 6' 6' 8' 25'			
Schorling-Sanford Achievement Test in Plane Geometry	1. Completing Sentences 2. Drawing Conclusions from Given Data 3. Judging Correctness of Conclusions 4. Analyzing Constructions 5. Computation Total	8' 12' 10' 10' 12' 52'			
Seashore Measures of Musical Talent	1. Pitch Discrimination 2. Intensity 3. Time 4. Consonance 5. Tonal Memory	4'-5' 4'-5' 4'-5' 4'-5' 4'-5'	.70 .62 .59 .34 .74	.31 .33 .33 .32 .29	100 100 100 100 100

(c) As given by Schorling.

TABLE 1 (Continued)

NAME OF TEST	UNIT ABILITIES COVERED	TIME LIMITS	RELIABILITY COEFFICIENTS	P.E. $\frac{\infty-1}{S.D.}$	N
Van Wagenen English Composition Scales	1. Thought 2. Content 3. Structure 4. Mechanics	None	.55 .67 .70 .50	.33 .32 .31 .34	50 50 50 50
Wakefield Diagnostic English Test	1. Noun Constructions (2 parts) 2. Verb Constructions (2 parts) 3. Voice (2 parts) 4. Mood (2 parts) 5. Tense (2 parts) 6. Classification of Sentences 7. Kinds of Clauses 8. Sentence Structure (2 parts) Total			.82	80
White Latin Test	1. Vocabulary 2. Translation of Sentences Total	15' 20' 35'	.38	.33	67

permit of more or less diagnosis. Some tests which claim diagnostic powers have been omitted in the light of the above criteria, and a few not claimed by their authors to have diagnostic value are included because they do permit of analysis of pupil strengths and weaknesses.

Of the seven requirements listed, the third is the one best met by the tests of Table 1. Requirement 1 is met to some extent by all the tests mentioned. However, there are seldom more than a half-dozen "units" covered by any test in the table. Requirement 5 is met in all cases fairly well, but it should be pointed out that no tests requiring tabulations of specific errors are included. The greatest shortcomings of diagnostic tests today center about requirements 2, 4, 6, and 7. Requirements 2 and 4 merge in the practical situation and reduce to the question whether each unit of the test which is to yield a score for an individual pupil represents a wide enough sampling to be a stable and reliable statistical measure. Chapters V to XIII in Part II of this volume present a considerable amount of evidence on the reliability coefficients of most of the tests of Table 1. Study of the data of these chapters shows that it is unusual for the reliability coefficients of high school tests to rise above 0.85 to 0.90 for the *total scores* made by typical classes. If, then, the pupil receives separate scores on from two to six or more sub-units, it needs no proof to show that the unit scores are not reliable enough to stand alone as a measure of an individual pupil. (This does not mean, however, that *class* diagnosis is impossible, since the averaging of twenty or thirty scores does give a stable and reliable measure of the *class*.) It is probably true that few of the units yielding separate scores within these various tests have reliabilities above 0.50 for typical class groups, and hence would not have very great significance in picturing the strengths and weaknesses *within* the total subject so far as individual pupils are concerned.

Too great stress can hardly be placed on this point, since test users have not been aware in the past of the errors of interpretation which arise from unreliability of test scores. Chapter XX of Part IV should also be consulted in this connection. Requirements 6 and 7 point toward an obvious situation in the use of diagnostic tests. Unless the teacher can "translate" the results of testing into remedial teaching based upon the strengths and weaknesses revealed by the tests, no constructive results arise from the testing. If the units are sufficiently analyzed, if they are reliably measured, and if they parallel suitable teaching and corrective units in the textbook or course of study, remedial work based upon the test scores is possible. If not, the tests lose their greatest usefulness. It is unfortunate that so few so-called diagnostic tests go the whole way outlined by these seven requirements.

Table 1 unfortunately is not complete with respect to several of the tests listed. The writers lacked data on many of the tests, and search of the literature yielded no results. All the data on reliability have been computed from tests given, scored, and treated statistically under the direction of the writers.

Where two or more forms of a test are provided, the reliability coefficients represent correlations of one form against a second form. If the test has but one form, the method used, except for the Seashore tests, was the correlation of odd- and even-numbered items, the resulting r being "stepped up" by means of the Spearman-Brown prophecy formula (see Chapter XX, page 359). In the case of the Seashore tests, each r reported was the correlation of percentile equivalents of the first testing with those of a repeated testing after some weeks' intermission.

The pupils used were presumably typical high school classes, except in the case of the Seashore tests, for which normal school students were used. The numbers are hardly

adequate at times, but in most cases the errors of sampling are probably not large. The last column at the right gives the numbers of cases.

The second column, entitled "Unit Abilities Covered," was usually obtained from the titles of the tests or sections of the tests, the wording being supplied by the present authors only when otherwise not obtainable.

The second column from the last gives the ratio of the probable error of an estimated true score (P.E._{est.}) to the standard deviation of the distribution of obtained scores (S.D.). Usually, the average of the standard deviations of the distributions of the two forms of the test was used. For details of the computation and meaning of these ratios, see Chapter XX. For the present the following facts will be sufficient for the interpretation of such ratios:

Ratios from .10 to .19 are highly satisfactory, indicating a sufficiently reliable test score for individual diagnosis.

Ratios from .20 to .29 are fairly satisfactory, indicating that the tests have some value for individual diagnosis.

Ratios from .30 to .39 are not satisfactory for individual diagnosis, but the tests can be used with confidence for purposes of class diagnosis.

Where the reliability is quoted for the entire test, it should be remembered that the reliabilities of the various sections of the tests are probably considerably lower than the figure given for the test as a whole.

Examination of Table 1 offers convincing evidence that present tests have made but the smallest beginnings in their efforts to yield genuine diagnosis of individual pupils.

CHAPTER THREE

USES AND LIMITATIONS OF TESTS IN THE HIGH SCHOOL (*Continued*)

III. TESTS IN GRADING, PROMOTIONS, AND SECTIONING

The general problems of classification. The classification of pupils covers at least three types of school practices: (a) the assignment of pupils to the proper grades, (b) the division of grades or classes into sections, and (c) vocational guidance and assignment to proper courses of study.

The first of these, assignment of pupils to the proper grades, is chiefly an elementary school problem. In high school it arises at times in subjects like English, foreign languages, and mathematics which are continuous over a period of years, especially with pupils who have been transferred from other schools. There is also the constantly recurring problem of the pupil promoted to high school who is not yet ready for high school work. Both educational and mental tests may be of assistance in grading such pupils.

Tests in determining promotions. Educational tests are always valuable in supplementing school marks and teachers' judgments in determining promotion or non-promotion. The great freedom of the tests from personal biases, their greater reliability, and the availability of norms eliminate much of the uncertainty attending decisions about promotions based on the usual data.

Tests should be used to *supplement* rather than to replace the more usual methods of promotion. It might not be defensible to argue that all pupils should be given educational tests in all subjects as a part of the final examinations, but the statement that *all backward and doubtful cases should be tested before final decision is made relative to promotion* is not without a sound basis. When we remember that the aver-

age traditional final examination seldom is more reliable than .60 to .70, any sources of refinement of judgments based upon such examinations should not be ignored.

The question is often raised, after educational tests have been given, how the test scores can help decide *how many* pupils should fail. No answer is possible to this question. The faculty should agree upon a reasonable per cent of failures purely as a matter of defining what a failing or a passing mark means. If it is done without slavish adherence to numerical values, there is much to be said for the practice of defining failing work as that done, over a period of years, by the lowest 5 per cent, or 7 per cent, etc., of the class. Such a per cent, if adopted as a school practice, should be looked on purely in the light of definition, and with conscious regard to the fact that certain classes will at times have no deserved failures and that other classes might occasionally have twice the stated per cent.

It is at this point that the standard test is capable of real service. In the light of the norms of average accomplishment provided by the test it is possible to determine *which classes* should depart from the adopted grading plan, and in *which direction*. Experience with letter-grading schemes carrying stated percentages of pupils to be given each letter grade has shown that, without an outside point of reference (e.g., a standard test), teachers unconsciously tend to grade systematically too high or too low over a period of years.

One further point about promotions and failures should be borne in mind in connection with all test and examination practices. *The primary aim of all examinations, tests, or school marks is that of ranking the pupils in the approximately true order of merit upon a scale of ability.* No grading system, whether it be the traditional 100 per cent scale or the A, B, C, D scheme, can possess any greater refinement from the standpoint of measurement than is possessed by the under-

lying arrangement of the pupils in rank order of achievement. The rank orders are the fundamental measures; the per cents or letters finally assigned are at all times artificial labels having no claims to validity not inherent in the rank orders.

To summarize: the rank orders answer the question as to *who*, if any, are to fail; the school authorities agree as a matter of pure definition *how many* are to fail. It is true, however, that a thoroughly trustworthy educational test will often be of material aid in deciding such issues.

Sectioning of high school classes. The high school with an enrollment of 150 or more pupils usually must face the question of class sectioning. If there are 50 or more entering ninth-grade pupils, there will be two or more sections in English, mathematics, history, and probably science and foreign language. The older practice managed this sectioning chiefly as a matter of avoiding conflicts in pupils' schedules and permitting free choice on the part of pupils. With the growth of our knowledge about the range of individual differences it soon became evident that sectioning of classes offered an opportunity for segregating pupils into relatively homogeneous groups; i.e., groups of approximately equal learning capacity and progress rates.

One of the objections sometimes voiced against sectioning of classes is that this practice increases the number of classes to be taught. If we except the small high school where sectioning is not ordinarily done, it can be shown that sectioning will not increase, but on the contrary will often decrease, the number of classes. To illustrate, let us assume that there are 100 entering pupils in algebra. Due to the relative difficulty of this subject, classes are ordinarily limited to 25 pupils. If these 100 pupils are sectioned at random or for convenience in programing, four sections and four teachers will be needed. However, if these 100 pupils could be

given intelligence, aptitude, or prognosis tests, it might be possible to form three sections, somewhat as follows:

	SUGGESTED NUMBERS
(1) Fast-moving section	45
(2) Normal or average section	35
(3) Slow-moving section	20
	<u>100</u>

This plan of sectioning would eliminate one section and one recitation per day. Experience has shown that it is, on the whole, more satisfactory to teach 40 to 45 pupils with an IQ range of 110-125 or from 95-105 than 20 to 25 pupils ranging in IQ from 80 to 130.

Sectioning upon a basis of ability permits differentiation of the course of study and adaptation of teaching methods to care for individual differences. The wide disparity of raw material normally entering high school may be made more concrete by an actual illustration. Table 2 below shows the range of educational abilities found among 125 entering ninth-grade pupils. Ability is expressed in terms of educational age earned upon the Stanford Achievement Test, a battery of tests covering most of the important elementary school subjects.¹

TABLE 2

SHOWING THE RANGE OF INDIVIDUAL DIFFERENCES AMONG 125 NINTH-GRADE PUPILS EXAMINED BY MEANS OF THE STANFORD ACHIEVEMENT TEST

EA	12-6	13-0	13-6	14-0	14-6	15-0	15-6	16-0	16-6	17-0	17-6	18-0	18-6
%	1.6	3.2	7.2	5.6	9.6	15.2	16.8	10.4	13.6	8.0	4.0	4.0	0.8

The total range of abilities found was about six years of educational age, which is about six school grades. The possibility of sectioning such a group into four differentiated progress groups by means of an instrument such as the

¹ Manual of Directions for the Stanford Achievement Test (World Book Company, 1925), Table 14, page 59.

Stanford Achievement Test is evident. This test could be given on the first day of school, while classes are being organized, and the blanks could be scored within twenty-four hours with no real loss of class time. The sections could be completely formed by the second day of school.¹

The following assignment of these 125 pupils to four sections is given as a suggestion. There are many other possibilities, although the scheme presented is probably as defensible as any.

Section A	Slow section	12-6 to 14-0	22 pupils (17.6%)
Section B	Low average	14-6 to 15-0	31 pupils (24.8%)
Section C	High average	15-6 to 16-0	34 pupils (27.2%)
Section D	Superior	16-6 to 18-6	38 pupils (30.4%)

The slow section has purposely been kept small and the superior section allowed the largest number, partly as a matter of convenience in breaking the distribution into quarters but more especially because the graduation in size of sections, if any, should take place in the direction shown. Thirty-eight pupils in the superior sections may seem like a large number, but in the opinion of many the teacher of the slow section will still have the most difficult task in spite of the smaller numbers.

The techniques of classification in the high school. The problems of classification, sectioning, and guidance are so closely interrelated that it is worth while to present an outline of the place of educational and mental tests in this general field.

¹ In many schools such tests as the Stanford Achievement Test are regularly given in May or June at the close of the eighth-grade work, in which case the scores would already be available to the high school principal and teachers. The only additional testing which need be done in this case would be on the pupils entering from other school systems. The same situation would hold for sectioning by means of intelligence tests.

AN OUTLINE OF CLASSIFICATION METHODS

I. Classification upon entrance to high school

A. Sectioning of classes

1. Methods possible

- (a) Elementary school marks
- (b) General intelligence tests
- (c) Educational test batteries over elementary school work
- (d) Prognosis, aptitude, and placement tests

B. Vocational guidance and selection

1. Methods possible

- (a) Courses in vocational study and guidance
- (b) Student counseling
- (c) Prognosis, aptitude, and placement tests

II. Classification in the upper high school grades (in subjects continuing over a period of years with prerequisites)

1. Methods possible

- (a) Marks earned in introductory and prerequisite courses
- (b) General intelligence tests
- (c) Educational tests over introductory and prerequisite courses

The foregoing outline calls for four more or less distinct types of quantitative measurements: (1) school marks, (2) intelligence tests, (3) educational tests on introductory and basal subjects, and (4) special prognosis, aptitude, and placement tests.

The order of listing above is, in the opinion of the authors, the approximate order of increasing usefulness of the four measures from a theoretical point of view. Practical circumstances, and in many instances the non-availability of good tests of the third and fourth types, make such an

evaluation almost impossible. Of the relative inferiority of teachers' marks there can be little doubt, in view of the large amount of experimental evidence on this point.¹ The use of previously earned marks was the first method of classification to be adopted and has no theoretical objections. Practically, however, such marks seldom are more reliable than 0.60 to 0.75 and hence do not offer a very secure basis for classification. This unreliability has always been one of the strongest arguments for the use of the standard test. A second practical objection to the use of marks in classification arises from the great diversity of marking schemes in existence; e.g., the "excellent-good-fair," etc., plan, the "A-B-C," etc., system, the 100 per cent scale, and many others. Ordinarily there will be pupils in a given high school class from a great many different elementary schools, and it will be difficult to equate such varying sets of marks to a common basis.

Intelligence tests for classification. Intelligence tests have generally proved of great value in classification, particularly the Binet-Simon individual examination. The group tests of intelligence are somewhat less valuable, but this disadvantage is probably more than offset by the economy of group testing. The following outline sets forth some of the main advantages and limitations of intelligence tests for classification purposes :

Advantages

- (1) Economy of time, labor, and money cost.
- (2) Ordinarily a single test sitting is needed.
- (3) Scores can be used for classification in several different high school subjects.

¹ See Chapter XIV of Part III for a brief account of the unreliability of teachers' marks. For a fuller account, see Ruch, G. M., *The Improvement of the Written Examination* (Scott, Foresman & Co., 1925), especially Chapters I-III.

- (4) More reliable than teachers' marks, or even most educational tests.
- (5) Experiments have shown fairly high predictive values for such high school subjects as English, mathematics, science, languages, and history.

Limitations

- (1) Less direct in their action than educational measures; i.e., they measure underlying capacities rather than school abilities *per se*.
- (2) Are not very predictive of many high school subjects; e.g., manual arts, commercial subjects, music, drawing, spelling, etc.
- (3) There is more likely to be objection on the part of parents to classification by mental tests than to classification by more direct educational measures.

There are a number of serviceable intelligence tests which are standardized for use in high school classes. The following list, although not complete, shows the better-known group tests of general intelligence:

1. *Dearborn Group Intelligence Test, Series II*, for Grades 4-9. J. B. Lippincott Company.
2. *Haggerty Intelligence Examination, Delta 2*, for Grades 3-9. World Book Company.
3. *Illinois General Intelligence Scale*, for Grades 3-10. Public School Publishing Company.
4. *Miller Mental Ability Test*, for Grades 7-12. World Book Company.
5. *Myers Mental Measure*, for Grades 1 to College. Newson & Co.
6. *Otis Group Intelligence Scale, Advanced Examination*, for Grades 5 to College. World Book Company.
7. *Pressey Senior Classification Test*, for Grades 7-12. Public School Publishing Company.
8. *Terman Group Test of Mental Ability*, for Grades 7-12. World Book Co.
9. *Thorndike Intelligence Test for High School Graduates (and College Freshmen)*. Bureau of Educational Research, Teachers College, New York.
10. *Thurstone Psychological Examination* (High School Seniors and College Freshmen). Carnegie Institute of Technology, Pittsburgh, Pa.

Educational tests for classification. The third type of measures outlined for classification purposes on page 8 are educational test batteries covering the previous preparation of the pupils in subjects basal, contributory, or prerequisite to high school subjects. To choose an illustration, consider first-year algebra as a typical high school subject. The elementary school abilities basal to success in algebra may be analyzed offhand as follows:

- (a) General intelligence
- (b) Arithmetic computation
- (c) Problem-solving ability in arithmetic
- (d) Reading ability (particularly in algebraic problems)

If educational tests were to be used for classificatory purposes in first-year algebra, few will disagree that tests of the types (b), (c), and (d) above would probably be the most useful. In an investigation carried out under the direction of one of the authors, the following results were obtained.¹ In this investigation, ability in algebra was determined after nine months of study by the combined scores on the Hotz and Douglass Algebra tests.

CORRELATIONS OF ABILITY IN ALGEBRA WITH:

1. Arithmetic computation and reasoning (Tests 4 and 5 of the Stanford Achievement Test)	0.62
2. Reading (Tests 1, 2, and 3 of the Stanford Achievement Test)	0.44
3. Intelligence (Otis and Terman tests pooled)	0.49
4. Chronological or life age	- 0.22
5. Interest in mathematics (an objective test prepared for this investigation)	0.12

The arithmetic test proved to yield the highest prediction. (It should be pointed out that the five tests listed above were all given at the beginning of the year in advance of study of algebra; only the Hotz and Douglass tests were given at the

¹ McCoy, J. P., *An Analysis of Algebraic Abilities* (1924). Unpublished.

end of the year.) The intelligence test stood second in predictive value, and the reading test was third in order. The other two measures were shown to have slight value. In fact, when all five tests were properly weighted and combined (by a multiple regression equation), the prediction was only increased to 0.644, in comparison with 0.62 for the arithmetic test alone, thus proving that there would be little gain in giving both an intelligence test and an arithmetic test and that there would be little practical justification for more than the one test; viz., the Stanford Arithmetic Examination.

The main advantages and limitations of batteries of educational tests for classificatory purposes may be listed as follows:

Advantages

- (1) Educational test batteries are superior in reliability to teachers' marks.
- (2) They measure the prerequisite or underlying abilities in a more direct and specific fashion than do intelligence tests.
- (3) A battery of good educational tests may be made to show a reliability of from .90 to .98 for two or three class periods of testing.
- (4) Educational tests indirectly measure the same abilities as do the intelligence tests, since intelligence has controlled in part the abilities making for high or low scores on the educational tests.
- (5) Experiments have shown educational tests, used as here suggested, to have as high predictive value as have intelligence tests, or possibly even higher.
- (6) Since several educational tests are usually given as a battery, there is diagnostic value to such a battery in revealing strengths and weaknesses of individual pupils.

Limitations

- (1) A battery of educational tests will require more time than a single intelligence test.
- (2) The pupil cost may be somewhat greater in the case of intelligence tests.
- (3) For some school subjects there are no available tests of real merit.

The following list of educational test batteries suggests the more useful series of classification tests:

1. *Illinois Examination*. Public School Publishing Company.

Contents:

- (a) Reading test
- (b) Arithmetic test
- (c) Intelligence test

2. *Lippincott-Chapman Classroom Products Survey Test*. J. B. Lippincott Company.

Contents:

- (a) Arithmetic Fundamentals test
- (b) Arithmetic Problem test
- (c) Reading Selections test
- (d) Reading Continuous Passage test

3. *Pintner-Marshall Combined Mental-Educational Survey Tests*. College Book Company, Columbus, Ohio.

Contents:

- (a) a non-language mental test (6 tests)
- (b) a survey test covering reading, arithmetic, grammar, history, and geography

4. *Stanford Achievement Test*. World Book Company.

Contents:

- (a) Reading: Paragraph Meaning
- (b) Reading: Sentence Meaning
- (c) Reading: Word Meaning
- (d) Arithmetic: Computation
- (e) Arithmetic: Reasoning
- (f) Nature Study and Science
- (g) History and Literature
- (h) Language Usage
- (i) Spelling

The four tests listed above differ considerably in scope, content, reliability, and general usefulness.¹ They are all primarily intended for the measurement of the efficiency of elementary instruction. Their greatest usefulness will be realized if they are given at the end of the eighth grade so that the scores can be used not only in determining promotions to high school, but also for purposes of classification in the high school.

Prognosis and aptitude tests. The last type of measures to be discussed is the so-called aptitude and prognosis tests. These might be regarded as tests of "special intelligence" or special abilities underlying success in a given school subject. In devising such tests the effort is made to analyze the abilities making up such subjects as mathematics, English, foreign language, etc., into a number of constituent elements. Various types of analysis may be used in devising such tests; e.g., analysis into physiological capacities (as in the Seashore Tests for Musical Talent), statistical analysis (as in the Rogers Test for Diagnosing Mathematical Ability), or validation against a criterion of success in industry or business (as in the case of trade tests like those of Chapman, Thurstone, and others).

In theory, the aptitude test is the most promising of any which have been discussed here because of the directness with which such tests attack the matter of predicting future success. However, the development of such tests is to be regarded as in its infancy. The most extensive series of tests falling within the classification of prediction tests is the recent series known as the Iowa Placement Examinations, by Stoddard, Seashore, Ruch, and others.² These

¹ Ruch, G. M., "The Achievement Quotient Technique." *Journal of Educational Psychology*, Vol. XIV (1923), pages 334-343.

² Seashore, C. E., "College Placement Examinations." *School and Society*, Vol. XX, No. 515 (November 8, 1924), pages 575-577. Stoddard, George D., "Iowa Placement Examinations." *University of Iowa Studies in Education*, Vol. III, No. 2 (August 15, 1925), pages 1-103.

tests are properly classified as prognosis or aptitude tests and can be used in both high school and college, although they were developed more especially for the sectioning of freshman college students. These Placement Examinations are listed along with other tests of their type in Table 3.

All the tests listed in Table 3 have their avowed limitations, but all have demonstrated sufficiently their validity and usefulness to warrant trial. The Seashore Measures of Musical Talent are based upon years of investigation in the laboratory and represent the sole measures available for the study of musical aptitude. The Stenquist Mechanical Aptitude Tests have a high reliability (at least .90 for average classes) and have shown correlations as high as .84 with shop teachers' ratings of success. The Iowa Placement Examinations have gone through a preliminary trial edition and revision. During the year 1925-26 at least 75,000 college freshmen were sectioned by these tests. Thurstone has found his Vocational Guidance Tests to be more highly predictive of engineering success than are the high school records of engineering students. Data on the value of several of the other tests have been published by their respective authors.

Educational tests *vs.* mental tests for classification. In order to summarize some of the recommendations brought forward in this section, the following statements are given as a tentative summary of present knowledge:

1. Teachers' marks are the least promising of the four types of measures discussed in this chapter. Such marks are probably much more open to question from the standpoint of reliability than of validity.
2. In theory, at least, prognosis and aptitude tests are the best of the four types of measures because of the directness and analytic character of such tests. Since so few good tests of this type are available at present, prognosis tests will seldom be the choice of the high school administrator.

TABLE 3
A BRIEF LIST OF PROGNOSIS AND APTITUDE TESTS

SUBJECTS OR ABILITIES	NAME OF TEST	RANGE OF APPLICATION
Music	Seashore Measures of Musical Talent (Pitch, Intensity, Time, Consonance, Tonal Memory, etc.)	Children and adults generally
Mechanical Ability	Stenquist Mechanical Aptitude Tests	Grades 6 to 12
Mathematics	Rogers Test for Diagnosing Mathematical Ability I. P. E. ¹ Mathematics Training I. P. E. Mathematics Aptitude	Grade 9 H. S. and College H. S. and College
Languages	Wilkins Prognosis Test in Modern Languages I. P. E. For. Lang. Aptitude I. P. E. French Training I. P. E. Spanish Training	H. S. and College H. S. and College H. S. and College H. S. and College
Engineering	Thurstone Vocational Guidance Tests (Arithmetic, Algebra, Geometry, Physics, and Technical Information) Iowa Placement Examinations (Listed throughout this table)	H. S. Seniors and College Freshmen College Freshmen
Clerical	Thurstone Clerical Examination	Selection of office clerks
Typists	Thurstone Typist Examination	Typists
Chemistry	I. P. E. Chemistry Aptitude I. P. E. Chemistry Training	H. S. and College H. S. and College
Physics	I. P. E. Physics Aptitude I. P. E. Physics Training	H. S. and College H. S. and College
English	I. P. E. English Aptitude I. P. E. English Training	H. S. and College H. S. and College

¹ I. P. E. is used as an abbreviation for "Iowa Placement Examinations."

3. Where fairly sure offhand analysis will show a number of previously taught component abilities, a selected battery of tests covering these basic abilities will probably yield the highest prediction easily obtainable. Experience must decide the truth of this recommendation for individual school subjects.

4. In the absence of available experimental evidence, the safest general practice in school classification is probably the general intelligence test, at least for the more highly verbal or linguistic school subjects.

IV. USES OF TESTS FOR RESEARCH PURPOSES

Tests in school investigations. A considerable amount of investigation and research is now done by school systems through research bureaus, supervisors, and teachers. Although the methodology of school research is outside the scope of this volume, a few comments on the rôle of educational and mental tests in such investigations may be in place. At least two major types of school research call for the use of test methods in controlling the conditions of the experimentation; viz.:

- (a) Studies of the relative merits of two or more proposed methods of instruction, and
- (b) Studies dealing with such factors as the rates of learning of children, amounts of drill or practice required, the influence of mental maturity on learning, racial and sex differences in learning, etc.

Standard tests enter into such investigations in two principal ways; first, in establishing the equality of the two or more experimental groups at the outset of the experiment, and secondly, in measuring the end-product or attainments of the several experimental groups.

For example, if it is desired to study the relative merits of

the additive and subtractive methods of teaching subtraction, the procedure would be somewhat as follows:

1. A large group of children, just before beginning work on subtraction, is tested on its previous knowledge of arithmetic (chiefly addition, in this case). The test employed here might be the Courtis or Woody addition scales.

A mental test is given, preferably the Binet, or if not the Binet, some group intelligence test.

2. On the basis of both sets of test scores from (1) above, the pupils of the entire group are paired off so that two groups equal in mental ability and knowledge of arithmetic are formed.

3. One of the two groups is taught the additive method and the other the take-away or subtractive method.

4. After subtraction has been completely taught, a standard test on subtraction is given and the mastery attained by the two groups is compared.

5. The appropriate statistical methods are then applied to determine whether significant differences have been established.

Tests have played an important part in the development of psychological methods of teaching, both for normal and for atypical children. The hundreds of investigations on methods of teaching spelling, on increasing rate and comprehension in reading, on racial and sex differences in mental ability, etc., have been made possible through the development of educational and mental test methods.

V. TESTS IN THE MOTIVATION OF LEARNING

Motivation as a neglected aspect of testing. When educational measurement was in its beginnings, it was a common practice for the superintendent to keep the sole set of test records on file in his office. Gradually the practice grew

of making these records available to the classroom teacher as well. The next step, and an important one, is that of placing the test records before the child as an incentive to improvement.

Learners improve most rapidly when their successes and failures are known to them at the time of practice. This principle, which may be termed the "principle of learning with knowledge of results," is one of the most effectual motivators known to students of the psychology of learning, and the standard test record offers a crucial opportunity to capitalize on a motivating situation of great force.

The pupil should be allowed to keep his own record, preferably in the form of a graph, and to record each test score. It is wise to place the emphasis on the pupil bettering his own past records rather than on bettering the published norms, since there will always be backward pupils who cannot reasonably expect to surpass the test norm but who can better past records.

CHAPTER FOUR

CRITERIA FOR THE SELECTION OF EDUCATIONAL TESTS

General outline. In view of the fact that there are often a number of available tests for the same school subject, an outline of certain critical questions to be asked in selecting tests is presented below. It is to be expected that this outline will not fit all types of tests and scales equally well, since these differ greatly in aim and scope. For this reason no attempt is made to supply criteria covering mental tests.

CRITERIA FOR SELECTING TESTS

I. Validation

A. Curricular

1. Does the test parallel good teaching practice?
2. How has the social utility of the test content been guaranteed?
3. Does the test provide an adequate sampling of the important topics or divisions of the subject?

B. Statistical

1. What correlations against outside criteria have been computed?
2. Have the test items individually been submitted to experimental try-outs in order to determine:
 - (a) That no large percentage of the items are failed by all pupils?
 - (b) That no large percentage of the items are answered correctly by all pupils?
 - (c) That the items show consistent increase in the percentage of successes with successive age or grade levels?

II. Reliability

1. Has the reliability of the test been established experimentally for some standard group such as an unselected grade, age, or subject group?
2. How were the reliability coefficients obtained?
3. Are the measures of reliability expressed in some stable form (e.g., the probable error of a score) which is more or less independent of the range of talent used?
4. Is the test reliable enough for individual measurement, or should it be confined to the measurement of groups like classes, schools, or school systems?

III. Ease of Administration

1. How complete and simple is the Manual of Directions or Examiner's Guide?
2. Are the test conditions well controlled by the instructions given in the Manual?
3. Are the directions to the pupils clear, detailed, and comprehensive?
4. Are samples and "fore-exercises" supplied where needed?

IV. Objectivity of Scoring

1. Do the authors provide convenient scoring keys and stencils?
2. Are adequate directions for scoring and computing scores given?
3. Is the test purely objective, or do the scorers have to exercise judgment in accepting pupils' responses?
4. About how many tests can be scored per hour?

V. Interpretation of Results: Norms

1. What kinds of norms are provided?
2. Are the norms based upon the performances of a

large number of pupils, or are they derived arbitrarily?

3. Is the interpretation of the pupils' raw scores simple and direct?

VI. Diagnostic Value of the Test

1. Is the test a general or "survey" test, or is it genuinely diagnostic?
2. If a diagnostic test, what principle or principles underlie the construction of the test?
3. How many different functions (skills, abilities, or aspects of the subject) are separately analyzed and measured?
4. Does the analysis of the total subject into unit abilities follow teaching practices and teaching needs?
5. Is the diagnosis individual or class diagnosis?
6. Does the test demand tabulation of individual pupils' errors in order to secure a diagnosis?
7. Is a remedial or corrective program based on the results of the testing provided (or suggested) through the Manual, or otherwise?

VII. Number and Equivalence of Duplicate Forms

1. How many equivalent forms of the test are provided?
2. Are the various forms equivalent? I.e.,
 - (a) Do they have the same average difficulty?
 - (b) Do they show the same range and variability of scores for the same group of pupils?
 - (c) Do they represent homogeneous and non-duplicative samplings of the same function (ability or subject)?

VIII. Time Requirements, Cost, Mechanical Features, etc.

1. Is the time required for giving the test as small as is consistent with reliable measurement?
2. Are the tests attractively printed and interesting in appearance?
3. Are the test blanks free from distractions by way of norms, directions to the teacher, etc.?
4. Is the paper of good quality?
5. Is the cost in keeping with the amount, scope, and reliability of the results yielded?

It is interesting to note that, as this book was going to press, Dr. Arthur S. Otis published a scale for rating educational tests.¹ Otis's form for rating is reproduced on the opposite page, without his detailed directions for its use, since the present chapter supplies the essential information for the understanding of the scale.

I. VALIDATION

Curricular validation. This term has been chosen to designate all those aspects of the validation of a test or scale which are concerned with educational objectives in contrast with purely statistical refinements. It refers to the "goodness" of the test as a test of an educational product. In more formal language, *validity is the degree to which a test measures what it is claimed to measure*. The term "validity" really includes reliability as one of its important characteristics, since without reliability of results a test cannot be highly valid.

The questions asked in the foregoing outline serve to define validity in a general way, and the details of the experi-

¹ Test Service Bulletin No. 13, "Scale for Rating Tests" (World Book Company, 1926).

SCALE FOR RATING TESTS	NAMES OF TESTS				
Manual (5)					
Validity (15)					
Reliability (10)					
Reputation (5)					
Ease of Administration (Total 15)					
(a) Preparation (4)					
(b) Time limits (4)					
(c) Explanation needed (3)					
(d) Alternative forms (4)					
Ease of Scoring (Total 15)					
(a) Objectivity (10)					
(b) Time required (3)					
(c) Simplicity (2)					
Ease of Interpretation (Total 15)					
(a) Norms (5)					
(b) Directions for interpreting (4)					
(c) Class record (1)					
(d) Application of results (5)					
Convenient Packages (5)					
Typography and Makeup (5)					
Test Service (10)					
Total (100)					

mental steps in the validation of a test are taken up in Chapters XVII and XVIII of Part IV.

The content of a valid test will look like the content of a well-worked-out course of study in the same subject, except that the test will be a limited sampling of the units of the course. Validity is one of the aspects of a test which in part is open to inspection by study of the printed test. In more quantitative terms, increases or decreases in the score values of the test should imply a corresponding increase or decrease in the educational abilities measured by the test. If Pupil X earns 40 points and Pupil Y earns 50 points on Test M, we must be able to assume that Pupil Y has more of the ability covered by Test M than does Pupil X.

In the validation of tests like spelling, vocabulary, arithmetic, etc., the principle of social utility is often applied; i.e., each word or example included in the test must demonstrate its usefulness in life. This suggests one of the important methods in the validation of test materials; viz., utilization of the published studies on the uses of spelling, vocabulary, arithmetic, and other subjects in the business world, the professions, the trades, etc.

In addition to validation with reference to social utility, the other principal validation methods of the "curricular" type are: analysis of courses of study, analysis of textbooks, analysis of examination questions, pooled judgments of competent persons, correlations with school marks, etc. Chapter XVII of Part IV describes such methods in greater detail. The important thing for the one selecting valid tests is to keep in mind the question, "How does the test author guarantee that his test parallels the objectives and content of the best teaching practices?"

The statistical validation of the test covers the experimental answering of the questions asked under I B of the outline of "Criteria for Selecting Tests." The prospective

user of a test has the right to know whether the "dead timber" has been weeded out of the test. "Dead timber" means items which are functionless or non-discriminating because they are too easy or too difficult for the pupils. Items passed by 100 per cent or by 0 per cent of the pupils are worthless in a test. Good items must show steady increases in the numbers of pupils passing them as we go up the age or grade scale; thus, an item which behaves as shown below is too erratic to be included in a test.

Grades	VII	VIII	IX	X	XI	XII
Per cent of successes . . .	31	43	44	75	60	81

Validity is such a broad concept that little more can be done at this place than to suggest a few of the issues which must be kept consciously in mind as the user studies the claims put forth by the test's author. It may be well to turn to Chapters XVII and XVIII at this point and glance through the discussion given there, covering the detailed steps in the validation of a test.

II. RELIABILITY

Definition of reliability. "Reliability" refers to the *accuracy with which the test measures whatever it does measure*. This is not necessarily what the test claims to measure. To take a crude illustration, the Hotz Algebra Scales might be used to measure will power, a thing for which they were not intended. The reliability of the Hotz tests would not be impaired by such misuse, even though the tests were not *valid* for measuring will power. A cheap thermometer will be less reliable (less accurate) than a more expensive warranted article. Similarly, two Latin tests might measure the same ability, but the scores on the one might be more accurate (reliable) than on the other. This might be due to the

greater length (more adequate sampling) of the better test, to its greater freedom from personal opinion in scoring, to the fact that the items had been more carefully chosen by experiment, to the inclusion of less "dead timber," or to a combination of these or similar causes.

If a test is given and Pupil A earns a score of 75 points, we know that these 75 points represent some position on a more or less arbitrary scale of ability. We should not think that the score 75 is his *true* score. If the test were repeated, or if a second equivalent form (sampling) were given, we should not be surprised if his second score proved to be 71, or 77, or even 81. We should expect changes with repetition of the test within certain limits, which we can call the *probable error of the score* (if stated in certain quantitative terms). The greater the range of this expected fluctuation, the less reliable the test.

Reliability is, then, the stability of the obtained test scores with repeated equivalent samplings, and is a measure of the confidence which may be placed in the test as a measuring device. Reliability is most often stated in terms of *reliability coefficients*; i.e., correlations between the scores earned by a group of pupils on two equivalent forms of the test. The detailed explanation of the statistical determination of reliability is reserved for Chapter XX of Part IV.

Reliability depends upon many factors, a number of which can be summarized by the following statements:

- (1) Other things being equal, the longer the test (the greater the number of test items), the more reliable it is.
- (2) Reliability in a test is in part due to the degree of objectivity of the scoring (the freedom of the scoring from personal opinion).

- (3) Reliability is decreased by "catch" questions, faulty wording, poor sentence structure, inadequate directions or samples, distractions in the test conditions, nervousness on the part of the pupils, and a variety of similar conditions.
- (4) Some school subjects appear to behave less reliably in test form than others; e.g., grammar seems to be more refractory than vocabulary, for reasons not entirely understood.
- (5) Reliability is decreased by items which show erratic changes in the percentages of successes from grade to grade or age to age.
- (6) Some test items show greater "discriminatory power" than others; i.e., they distinguish smaller differences in pupil ability.

The importance of reliability. Reliability is second only to validity in evaluating the worth of a test. Indeed, reliability influences validity, since a test which does not yield accurate and stable numerical scores cannot be highly valid. However, any test, be it ever so reliable and valid, loses validity when misapplied, just as thermometer readings, no matter how accurate, would lose all validity if interpreted as measures of barometric pressure. The list of "criteria" at the beginning of this chapter suggests the questions to be asked in deciding the claims of a test to reliability. Such questions should be answered in an easily accessible place, preferably in the Manual of Directions.

The degree of reliability of the measurement afforded by a test is unfortunately very little open to determination by casual inspection of the printed test. Such information must be obtained by careful experimentation combined with the use of appropriate statistical computations. Such determinations are properly the task of the test maker, not of

the test user. Although the failure of the former to secure and publish data on reliability of his product does not imply that the test is not serviceable, the prospective user can rightly discriminate against such a test as a matter of conservatism. Most of the carefully constructed and validated tests have been carefully examined for reliability, and the data necessary for the evaluation of the test are presented along with the claims made for the test in the Manual or elsewhere.

Common sense will tell something about the probable reliability of a test. Thus, a 5-minute test in some complicated ability like Latin, geometry, or physics is practically certain to have a low reliability because such broad achievements cannot be measured accurately in five or ten minutes. On the other hand, a "narrow function" test like one on the mensuration of plane surfaces, or the vocabulary of first-semester Latin, or the laws of falling bodies, may sample rather adequately such a narrow field in ten to fifteen minutes if the test has been carefully refined during construction. It will help in our thinking about a test to remember that a topic or series of topics requiring from one to five or ten months to teach is not likely to be well measured in a 5-, 10-, or even 20-minute test. A distinction must be drawn here between the various types of tests. Survey tests for measuring entire schools or school systems can yield reliable measures with a very small number of items (10 to 50, perhaps), since we are here dealing with *average* scores of large numbers of pupils. At the same time, the score of any *individual pupil* might be entirely too unreliable to be taken at face value. In the case of diagnostic tests or other tests which provide for a number of separate scores (by parts, sections, separate abilities, etc.), each separate score, if it is to be used as a datum for interpretation, follows the general rules just given. This means that a diagnostic

test in problem solving in algebra which has five parts as follows —

- (a) Part 1. What Facts Are Given in the Problem?
- (b) Part 2. What Does the Problem Call For?
- (c) Part 3. What Operations Are Needed in the Solution?
- (d) Part 4. What Would Be a Reasonable Answer?
- (e) Part 5. Choosing the Correct Solution.

would have to demonstrate that each of the five parts (as well as the total score) is sufficiently reliable to be a trustworthy diagnostic measure. Such a test would have to be several times as long as a survey test yielding but one general score, both in terms of numbers of items and in working time.

What is a satisfactory degree of reliability? It is almost impossible to answer this question except in a rough way. Reliability coefficients are correlations, and hence their magnitudes are influenced by the range of abilities present in the group of pupils on which the correlations are computed. Other things being equal, a single grade group will yield a smaller reliability coefficient than two successive grade groups pooled. Three successive grade groups will produce a yet larger coefficient. A test might have a reliability of .40 or .50 in a fifth-grade class alone and show a coefficient of .90 to .95 for eight or ten grades pooled. Chapter XX of Part IV will show a more satisfactory expression for reliability than the mere coefficients of reliability.

In spite of the limitations imposed by the above facts, and certain other statistical considerations, some rough guides to thinking about reliability coefficients are given below. These figures are suggestive only and imply that the reliability has been computed on "average" or "typical" classes of a sufficient size to provide a stable sample.

RELIABILITY COEFFICIENT	INTERPRETATION OR SIGNIFICANCE
0.95 to 0.99	Very high; rarely found among present tests
0.90 to 0.94	High; equaled by a few of the best tests
0.80 to 0.89	Fairly high; fairly adequate for individual measurement
0.70 to 0.79	Rather low; adequate for group measurement but not very satisfactory for individual measurement
Below 0.70	Low; entirely inadequate for individual measurement although useful for group averages and school surveys

The authors hesitated to make the above statements, but it would seem that concrete criteria of some kind are due the reader. Part II of this volume presents data on the reliability of a large number of high school tests. It will be noted that the great majority fall in the lowest three groups. Justification of the above characterizations will be deferred for treatment in Chapter XX of Part IV.

III. EASE OF ADMINISTRATION

Adequacy of instructions. Ease of administration should be judged from two points of view; first and most important, the clarity of instructions to the pupil, and second, the clarity of the instructions to the examiner. Directions to the pupils are properly printed on the test booklets. If the test is broken into parts or sections, each section should be preceded by the directions for that unit, together with samples showing the pupil how he is to indicate his answers. The instructions to pupils should be full and very simple in phraseology. It must not be assumed that all pupils will hold in mind long and complicated directions. If the instructions are necessarily somewhat involved, numerous samples, "warming-up exercises," or "fore-exercises" should be provided.

The instructions to the examiner ordinarily *should not be printed on the pupils' test blanks*. This rule is often not observed, usually to the distinct disadvantage of the test. The argument for placing both sets of instructions on the test blanks is that it saves the trouble and expense of a Manual of Directions. A properly executed Manual of Directions should provide a vast amount of detailed information which, if supplied adequately, could not be placed on the test blanks for reasons of space.

Some of the objections to printing instructions for the examiner on the test blanks are: (a) The language in which such instructions are written may puzzle the pupils, (b) they are a source of distraction to the pupils, (c) many of the facts which the examiner should know are unnecessary for the pupil and even undesirable, (d) the test pages are likely to be made crowded and unattractive to the pupils, and (e) the pupil may become confused by the complexity of a double set of directions and miss the directions which explain the method of taking the test, thus losing time or otherwise penalizing himself. The same objections apply to placing the norms on the pupils' blanks or booklets.

There is a very simple rule which covers all these practices, one which is observed in good textbook construction and other instructional materials; viz., *Place on the materials intended for the pupil, only those printed directions which apply to him, reserving all instructions to the teacher or examiner to a separate place, preferably the Manual of Directions or Examiner's Guide*.

When one considers the multiplicity of details necessary for the examiner (e.g., accounts of the validation, reliability, norms, scoring, interpretation, etc.), it is evident that the pupils' blanks do not permit satisfactory treatment of such details.

A test whose administration cannot be learned by the

average teacher in an hour or two is not likely to succeed. At the same time the beginning examiner must be warned about a multiplicity of small but important details, such as instructions covering the distribution of blanks, the filling in of pupil information data, the observance of time limits, the breaking of pencils, the prevention of disturbing factors, and other requirements of good test conditions.

The questions listed under Item III in the outline of "Criteria for Selecting Tests" at the beginning of this chapter suggest critical points which may well be raised in choosing worth-while tests.

IV. OBJECTIVITY OF SCORING

Objectivity an essential in good tests. The historical development of educational measurement proves the importance of freeing the test from subjective factors in marking the papers. One of the main points of superiority of standard tests over traditional examinations is the elimination of the unavoidable errors of judgment as to the merits of answers. To this end most standard tests have adopted mechanical features of scoring, such as the use of transparent celluloid, perforated sheets, or printed strips which can be superimposed on the pupils' booklets, permitting comparison of their answers with those printed on the scoring stencil. For the same reasons standard tests have been arranged so that pupils underline, check, cross out, or otherwise indicate the correct answers, the amount of writing which the pupils must do being kept to a minimum.

The care with which the mechanics of answering and scoring of test items has been worked out is an important element in the selection of tests. Experience has shown that a test which is hard to score usually fails to secure wide usage. There are two aspects to the question of ease of scoring;

first, economy of time and effort, and, second, the more mechanical the scoring the more objective the results. Objectivity in turn reacts favorably upon the reliability of the test, and, less directly, upon its validity. No small part of the superiority of the rank and file of educational tests over the traditional school examination is to be explained by the greater freedom of educational tests from subjective factors in scoring.

The Manual for the test should contain a section on the method of scoring. Explanations should be given covering the handling of such matters as erasures, alterations, corrections for chance, illegibility of responses, attempts of pupils to "beat the game" by marking everything on the page (as often happens in multiple-response tests), etc. In tests of the true-false or yes-no type, pupils will often mark every one true or every one false in the attempt to get part of the items right. Rules must be given for treating this and many other similar situations.

The speed with which the tests can be scored is also an important item within certain limits. Keeping in mind the amount of scoring which ordinarily must be done, there is no great disadvantage attaching to a test requiring three minutes for scoring over one requiring one minute. If a single test requires ten or more minutes for correction, the question should be raised whether the results justify the time expenditure. The time factor must also be evaluated with reference to the length and general usefulness of the test. A brief survey test which can be corrected at the rate of sixty an hour might be much less worth while than a good diagnostic test scorable at a rate of but ten or fifteen an hour.

V. INTERPRETATION OF RESULTS: NORMS

Accuracy of norms. The evaluation of the accuracy of the norms provided with a test is ordinarily not very simple. In selecting tests, much must be trusted to the authors' accounts of the method of arriving at the standards. The numbers of pupils used in computing norms should be stated, although *numbers alone* guarantee little or nothing about the accuracy of norms. A test with norms based upon one thousand cases may very possibly represent typical or normal conditions better than one with five thousand or even ten thousand cases. The critical question here is, "How were the pupils providing the norms selected?" A few thousand pupils rigidly selected as a representative sampling are sufficient, as will be shown in Chapter XIX of Part IV. Too often test norms are based upon hundreds of thousands of records which have been voluntarily returned to the author of the test by test users. Such norms are almost certain to be biased or selected, since only certain types of schools (presumably the more progressive) use standard tests. Again, schools making a superior showing on the test are much more likely to send in returns than schools scoring low. In the third place, city schools are usually represented in such returns to an extent which "swamps" the small town and rural returns. The prospective buyer of a test should be much more concerned with the "how" of the norms than with the "number."

A norm is usually an average, a median, some percentile, a measure of variability, or some other measure derived from one of these. It is most often a mean or a median. Selecting a case where the norms are given by grades, let us assume the following figures:

GRADE	CASE I		CASE II	
	MEAN (Average)	NUMBER	MEAN (Average)	NUMBER
IX	60.3	2,000	60.3	25,000
X	71.8	2,000	71.8	25,000
XI	83.5	2,000	83.5	25,000
XII	91.2	2,000	91.2	25,000

The accuracy of such a mean is dependent upon two things: (a) the variability or spread of the distribution, and (b) the number of cases. Variability is usually expressed by the standard deviation (σ). The probable error of the mean can be computed from the formula,

$$P.E._M = \frac{.6745 \sigma}{\sqrt{N}}.$$

In the two situations listed above (Case I and Case II) the standard deviations (σ) would be approximately the same. Let us assume σ equal to 15.5 in both cases. The probable errors of the IX grade mean (norm) for both cases are, therefore:

$$\begin{array}{ll} \text{CASE I} & \text{CASE II} \\ \frac{.6745 \cdot 15.5}{\sqrt{2000}} = .23 & \frac{.6745 \cdot 15.5}{\sqrt{25000}} = .066 \end{array}$$

The probable error of the mean (60.3) for Grade IX for Case I is less than one fourth of a score point; for Case II it is about one fifteenth of a score point. Taking Case I as our point of reference, the following statements are approximately true:

- (1) The chances are even that the true value of the Grade IX mean lies within ± 1 P.E. on either side of 60.3; i.e., between 60.3 minus .23 and 60.3 plus .23 (between 60.07 and 60.53).
- (2) The chances are about 4 to 1 that the true value of the IX grade mean lies within ± 2 P.E.; i.e., between

60.3 minus $2 \times .23$ and 60.3 plus $2 \times .23$ (between 59.84 and 60.76).

(3) The chances are about 20 to 1 that the true value of the IX grade mean lies within ± 3 P.E.; i.e., between 60.3 minus $3 \times .23$ and 60.3 plus $3 \times .23$ (between 59.61 and 60.99).

By similar calculations it might be shown that the chance of any mean (norm) under Case I being in error by 1.0 score point is about 1 in 300. For Case II the chances are of course even less. The practical importance of the foregoing discussion reduces to the fact that even a few thousand cases establish norms with a high degree of accuracy, *provided the sampling is a random and unselected one*. If the ninth-grade norm under Case I is no more likely to be in error by 1 score point than about 1 chance in 300, and the difference between the norms for successive grades is about 10 points (as shown), there is small likelihood of any pupil's score being badly misinterpreted through errors in the norm. The real source of danger is in the error present in the pupil's score, as we have shown in the discussion under the section on reliability in this chapter. Attention is once more drawn to the fact that it is the *randomness* and *representativeness* of the cases supplying the norms, not the numbers, provided there are at least a few hundreds or a few thousands of cases.

Kinds of norms provided. In high school subjects, age and grade norms are less useful than is true in elementary school subjects. The three most common types of norms supplied with high school tests are:

- (a) Percentiles
- (b) Subject norms
- (c) *T*-scores¹ (or other measures based upon the standard deviation or other measures of variability)

¹ See page 350.

Subject norms usually provide only a single norm value, often by grades in subjects continuing over several years in high school. These have high reliability but are relatively less useful because exact estimates of the superiority or inferiority of a given pupil's score cannot be made. Such a norm shows the pupil to be above or below standard, but little more.

Percentile norms are not quite as accurate in general as *T*-scores or other measures based upon the standard deviation but are possibly more meaningful to the average teacher, due to the ease of thinking about percentiles. The median and the quartiles are but special names for the fiftieth, twenty-fifth, and seventy-fifth percentiles.

The simplicity of percentile norms is shown by the following norms for the Ruch-Cossmann Biology Test (revised to January 1, 1926) :

10% of the pupils reach or exceed	63.7
20% of the pupils reach or exceed	55.5
25% of the pupils reach or exceed	53.0 (Upper quartile)
30% of the pupils reach or exceed	49.8
40% of the pupils reach or exceed	44.2
50% of the pupils reach or exceed	39.6 (Median)
60% of the pupils reach or exceed	36.1
70% of the pupils reach or exceed	32.6
75% of the pupils reach or exceed	30.2 (Lower quartile)
80% of the pupils reach or exceed	28.2
90% of the pupils reach or exceed	23.1
Numbers	753

The pupil's accomplishment can be interpreted directly as falling into successive tenths of a distribution of typical pupils or into quarters (by use of the quartiles).

The *T*-score is similar to the percentile for practical purposes of understanding, the raw scores of the pupils being turned into values which are fractions of the standard deviation of the distribution entering into the norms. Tables are provided so that the *T*-scores may be looked up without

computation. For the details of the meaning and calculation of *T*-scores the reader is referred to Chapter XIX of Part IV.

The major issues in examining the norms of a test are the representativeness of the sampling and the ease of interpreting the raw scores, whether by percentiles, *T*-scores, or other similar measures. Convenient tables of norms should be provided.

VI. DIAGNOSTIC VALUE OF THE TEST

The evaluation of the diagnostic functions of a test. Section II of Chapter II has covered the salient points of the requirements for genuine diagnosis. There are so few really diagnostic tests in the high school field that there is little which can be added to the discussion in Chapter II. The questions listed in the "Criteria for Selecting Tests" should be clear both as to their meaning and their importance in the light of what has already been said.

Too much emphasis cannot be placed upon the fact that real diagnosis is time-consuming and more costly in terms of money than mere measurement of general proficiency without detailed analysis and separate measurement of the component abilities making up a school subject.

The available diagnostic tests suffer more from critical evaluation in the light of the criteria set up here than do the survey types of test, because of the greater difficulties attending the securing of adequate subject analysis, reliable individual scores, etc., and because of the difficulty of keeping the testing within practicable time and cost limits. Nevertheless, good diagnostic tests are worth the effort, since they lead to teaching reforms and remedial programs not nearly so easily possible through general survey tests intended to give a single "blanket" measure of a pupil or class. Present diagnostic tests in high school subjects are

not without value, but at the same time they are not "fool proof" and they should be interpreted with full recognition of their limitations.

VII. NUMBER AND EQUIVALENCE OF DUPLICATE FORMS

The need for duplicate forms. In order to provide for repeated application of the same test, two or more forms of the test are highly desirable. Two forms will serve almost all practical purposes. Two forms will allow testing at the end of each semester of the school year without repetition of the same test items. If the testing is done but once a year, the forms may be alternated by years. This prevents coaching and undue memory effects where the same test is repeated. Most of the best high school tests, fortunately, are standardized in two forms. A few have three or more. In selecting tests, it is reasonable that some weight be given to the existence of duplicate forms.

Equivalence of forms. The word "equivalent" as used in educational measurement has a technical definition. The main assumptions and conditions to be met by exactly equivalent forms may be stated:

1. The test items represented by the several forms of the test should be random samplings of a larger amount of valid and homogeneous material, which collectively covers the entire subject matter in a thorough manner.
2. There should be no duplication of items from form to form.
3. The average difficulties of the forms should be equal; i.e., it should be a matter of indifference which form is used.
4. The various forms should show the same spread of scores for a given lot of pupils; i.e., the standard deviations and other measures of variability should be, within reasonable limits, the same on all forms.

5. The scores of individual pupils should vary as little as possible from form to form; i.e., each form should be made long enough to provide stable and reliable individual measures.

The proof that these five conditions have been met satisfactorily should be furnished with the test, the Manual being the logical place to present such experimental evidence. Since nothing can be told about the equivalence of test forms by mere inspection, the test authors should supply all such information to prospective users of the product.

VIII. TIME REQUIREMENTS, COST, MECHANICAL FEATURES, ETC.

Time requirements. No rules can be laid down except to suggest again that abilities worth teaching for five to nine months are worthy of more than ten or fifteen minutes' time in their attempted measurement. Other things equal, the longer the time devoted to the actual testing, the better the measurement resulting. It is true that a test can be made very wasteful of pupils' time, but this seldom happens as a matter of fact. The real danger is from short and unreliable tests. The fear that more time will be consumed than is justifiable is largely unfounded. As a matter of historical precedence and development, short tests have been greatly in demand; but the insistence on brevity is certain to be outgrown in favor of accuracy of result. Many tests of widespread popularity date in their origin from long before the time when the concept of reliability of measurement was consciously in the minds of test makers, with the result that the most "economical" tests are not always the best selections. Much has been learned about test construction within the past decade, and much of this experience has thrown great doubt on the possibility of genuine measurement without paying the price.

As has been stated before, survey tests for group measurement need comparatively less time (and length in terms of items) than tests yielding single measures of individual pupils, and these in turn require less time than diagnostic tests which not only measure individuals but also many aspects of the accomplishment of individuals. A class period twice a year is little enough time for tests of general accomplishment of individual pupils. For diagnosis, if it is to be a serviceable diagnosis, four or five class periods a year are easily justifiable. Within reasonable limits the time factor can well be eliminated from serious consideration. The burden of justification falls at the door of the brief test, not the thorough one.

Cost. Educational tests are sold on a competitive market like safety razors, tooth paste, and groceries. The retail price reflects a variety of production costs: experimentation, printing, editing, advertising, selling, etc. All these factors vary with different tests. The cost of the original experimentation is by far the greatest single factor in a well-constructed test. To the certain knowledge of the authors, the range of actual experimental costs among existing educational and mental tests varies from less than one hundred dollars to more than ten thousand dollars. Other production costs probably show similar variations. Educational tests of some sorts can be sold at a profit at one cent a test blank, while others would be sold at a loss at five cents. No thinking man expects five hundred dollars spent for a popular four-cylinder car to buy the same absolute value as five thousand dollars applied on the purchase of a high-powered custom-built limousine. Tests, after all, are commercial commodities like textbooks, schoolhouses, teachers' salaries, and a hundred other items of school cost.

The question to be asked about a prospective test is not "How much does it cost?" but "Is it worth the price?"

In the main, schools will find that the purchaser gets about what he pays for in planning expenditures for the testing program.

Mechanical features. To some extent the mechanical features of a test influence cost, but not very greatly. The difference in the cost per test blank between good and poor printing or between good and poor paper is usually a very small fraction of the total cost. At the same time, the quality of such mechanical features may influence greatly the attitude of pupils toward the test. The rules which govern good textbook printing are generally adequate and applicable to tests. The uninteresting, scrambled, hodge-podge test will tend to evoke a corresponding attitude in the pupil. Too little attention has been paid by test publishers to the psychological effects of tests on pupils. Experienced test examiners will bear witness to the reality of an issue here.

Most of the objectionable mechanical features in certain tests arise from a dangerous practice of attempting to make the test booklets serve also as a Manual of Directions for the examiner. This practice can only result in confusion to the pupil, slighting of the needs of the examiner, and the omission of much valuable data on the worth of the test. It is to be hoped that needed reforms in this direction will be demanded by test users in the future.

*

PART TWO

**DESCRIPTIONS OF HIGH SCHOOL TESTS
BY SUBJECTS**

*

CHAPTER FIVE

MATHEMATICS

Introduction. Measuring the results of teaching high school mathematics is complicated by many factors. As with all high school subjects, there is a lack of agreement in educational objectives. Just what purposes are algebra and geometry expected to serve in the life of an individual student? What weight should be attached to mastery of the mechanics of mathematical procedure, what types of reasoning problems are desirable, and what transfer is expected to other subjects? Answers are based largely on opinion, and their diversity is reflected in the differences found in courses of study and in textbooks. However, the 1923 report of the National Committee on Mathematical Requirements¹ may be expected to render valuable service in clarifying and coördinating the content, methods, and aims of secondary mathematics.

The tests and scales for measuring high school mathematics, all developed within the last ten years, have simply taken the situation as they found it. The tests are fairly adequate for measuring the fundamentals, the machinery involved, although agreement is not reached even here. The broader problems of measuring the effects of mathematics on life situations and the determination of the educational aims of mathematics have scarcely been touched.

The value of existing tests in prognosis of pupil success is supplementary. Tests of achievement in mathematics are useful in guidance only when considered in connection with general intelligence, performance in other subjects, and certain personal characteristics.

¹ *The Reorganization of Mathematics in Secondary Education.* A Report by the National Committee on Mathematical Requirements under the Auspices of the Mathematical Association of America, Inc. (Published by the Association, 1923; 652 pages.)

The recent experimental findings of Schreiber¹ are of interest in this connection. His chief conclusions are:

1. With intelligence constant, the arithmetic abilities in addition and multiplication had little to do with success in algebra.
2. The chief criterion of success in first-year algebra is "ability to manipulate the machinery of algebra."
3. General intelligence is a substantial factor of success in first-year algebra.

The need for new measures in high school mathematics will grow with increased knowledge of the social and educational needs of (a) pupils who leave high school to enter various vocations, and (b) pupils who take higher mathematics in college. It is encouraging to know that the National Committee² found that "there appears to be no real conflict of interest between those students who ultimately go to college and those who do not, so far as mathematics is concerned." The question is, How much of the recommended two or two and one-half years of secondary mathematics is important in the activities of pupils who do not enter college?

I. ALGEBRA

Hotz First-Year Algebra Scales

Description of the scales. The Hotz Scales are designed to measure comprehensively the ability of pupils in first-year algebra. The following five scales are published:

1. Addition and Subtraction	4. Graphs
2. Multiplication and Division	5. Problems
3. Equations and Formulas	

The first two scales include fractions and radicals; the second two cover the chief instruments of quantitative think-

¹ Schreiber, Edwin W., "A Study of the Factors of Success in First-Year Algebra." *The Mathematics Teacher*, Vol. XVIII, No. 3 (March, 1925), pages 141-163.

² *Op. cit.*, page 43.

ing. The problems in the fifth scale are typical of first-year algebra courses. Two series of all scales are available. Series B is the longer, comprising from eleven to twenty-five items in each test, and requiring 3 hours and 20 minutes' actual working time to administer. Series A is about one half the length of Series B, is composed of exercises drawn from the latter scale, and requires 1 hour and 50 minutes' working time. It covers the same range of difficulty as the longer set. The intervals between successive exercises in Series A are approximately equal. If all five scales are given, Series A is recommended. If economy of time necessitates the selection of one or two scales from the entire series, Hotz recommends the Equation and Formula Scale, as it is the most comprehensive. The Problem Scale comes next in importance. These two scales can both be given in one class-hour. Each scale is printed on a separate folder, with sufficient space for the pupil's computations.

What the scales measure. The Hotz Algebra Scales begin with items which practically all the pupils can solve, and proceed with regular increases in difficulty. Only a relatively small proportion of any class can solve the last problems in each scale. Hence the factor of speed is minimized. Hotz's method of selecting the material for the scales has led to the inclusion of only such items as are generally taught in all first-year algebra courses. He "firmly believes that no essential algebraic process has been ignored." Series B is much more reliable for individual diagnosis, since it offers a wider variety of exercises.

Administration and scoring. Complete directions for giving and scoring are contained in the Teacher's Manual. The directions are simple and may be grasped readily by the pupils. Scoring is completely objective. Each answer is marked either right or wrong, and no partial credits are allowed. The scales can be given after three months of algebra

instruction, and thereafter according to the convenience of the teacher. Duplicate scales are not available, but the teacher may reduce practice effects by alternating the two series.

Interpretation and utilization of results. The uses of the scales may be stated as :

- (1) The measurement of class and pupil achievement;
- (2) The measurement of class and pupil progress; and
- (3) The diagnosis of difficulties.

The median score should be compared to the standard medians. A median too low shows that something is wrong, although the trouble may not be due to poor instruction. The textbook used, the length of the class period, the emphasis on different phases of the work, and the intelligence of the pupils are a few of the facts which must be taken into account in interpreting results. Progress is measured by comparing performance of pupils on the same scale or similar scales at various periods. In diagnosing pupil difficulties, it should be remembered that the important thing is not how many exercises the pupil can do in a certain amount of time, but what point in the scale he reaches. This is the distinguishing feature of a *scale*.

Validity and reliability. Exercises were placed in their relative positions on a projected linear scale on the basis of difficulty of solving. The P.E. was chosen as a unit. Hotz's Manual describes the method in detail. Reliability data are given in Table 4.

TABLE 4

DATA ON RELIABILITY¹ OF HOTZ FIRST-YEAR ALGEBRA SCALES, SERIES A

r	N	S.D.	P.E. score	P.E. $\infty.1$	P.E. $\infty.1$ S.D.	P.E. $\infty.1$ S.D.	NATURE OF GROUP
.92	175	8.70	1.7	1.6	.20	.18	Grade IX pupils

¹ For the meaning of the statistical computations in this and following tables, see Chapter XX of Part IV.

Norms. A number of standards are given in the Teacher's Manual. They comprise median scores on the various scales in both series, covering algebra instruction over the following periods: 3, 6, 8, 9, 10, and 14 months. Additional norms, describing more specifically the nature of the group, are given in Tables 5 and 6.

TABLE 5¹

GRADE NORMS FOR HOTZ'S FIRST-YEAR ALGEBRA SCALES, SERIES A,
MAY TESTING

	GRADE IX	GRADE X
<i>Addition and Subtraction:</i>		
Number of pupils	561	390
25-percentile	5.2	5.8
Median	6.9	7.3
75-percentile	9.1	8.7
<i>Multiplication and Division:</i>		
Number of pupils	570	388
25-percentile	5.7	5.9
Median	7.2	7.4
75-percentile	8.4	8.7
<i>Equation and Formula:</i>		
Number of pupils	478	385
25-percentile	6.2	6.7
Median	7.7	7.9
75-percentile	9.7	9.1
<i>Problems:</i>		
Number of pupils	566	394
25-percentile	4.5	3.9
Median	6.4	5.0
75-percentile	8.6	6.3
<i>Graphs:</i>		
Number of pupils	121	413
25-percentile	5.2	4.1
Median	6.2	5.0
75-percentile	7.0	6.0

¹ "Report of Division of Educational Tests for 1919-1920." *University of Illinois Bulletin*, Vol. XVIII, No. 21 (January 24, 1921), page 31.

TABLE 6¹COMPARATIVE RESULTS OBTAINED WITH HOTZ FIRST-YEAR ALGEBRA
SCALES (EELLS)

NAME OF SCALE	NUMBER OF SCHOOLS	NUMBER OF PUPILS
Addition and Subtraction	64	4168
Equation and Formula	61	4195
Multiplication and Division	39	2207
Problems	38	2233
Graphs	20	801

A. Median scores (all schools).

	ADDITION AND SUB- TRACTION SERIES A 12 WEEKS	EQUA- TION AND FORMULA SERIES A 12 WEEKS	MULTI- PLICA- TION AND DIVISION SERIES A 24 WEEKS	PROB- LEMS SERIES B 24 WEEKS	GRAPHS SERIES A 24 WEEKS
Hotz Standards	5.0	4.9	6.3	6.5	3.7
Eells	5.4	5.6	7.0	7.0	5.2
Approximate Difference in Months	2-3	1	1 $\frac{1}{4}$	1 $\frac{1}{2}$	2

B. Summary by Size of School (Eells); medians.

(Enrollment over 300 = Large; 100-300 = Medium; under 100 = Small)

CLASS OF SCHOOL	NUMBER OF SCHOOLS	ADDITION AND SUB- TRACTION SERIES A 12 WEEKS	EQUATION AND FORMULA SERIES A 12 WEEKS	MULTIPLICA- TION AND DIVISION SERIES A 24 WEEKS	PROBLEMS SERIES B 24 WEEKS	GRAPHS SERIES A 24 WEEKS
Large	16	5.6	5.8	7.3	7.6	5.8
Medium	21	5.5	5.7	7.2	6.7	5.9
Small	28	5.2	5.3	6.6	6.7	4.9

¹ Eells, W. C., "Hotz Algebra Scales in the Pacific Northwest." *The Mathematics Teacher*, Vol. XVIII (November, 1925), pages 418-427.

TABLE 6 (*Continued*)

C. Summary by States (Eells); medians.

NAME OF STATE	NUMBER OF SCHOOLS	ADDITION AND SUBTRACTION SERIES A 12 WEEKS	EQUATION AND FORMULA SERIES A 12 WEEKS	MULTIPLICATION AND DIVISION SERIES A 24 WEEKS	PROBLEMS SERIES B 24 WEEKS	GRAPHS SERIES A 24 WEEKS
Washington	33	5.3	5.5	7.0	7.1	5.3
Idaho	16	5.3	5.5	6.4	6.7	4.9
Montana	10	6.0	5.8	8.3	7.7	5.7
Oregon	5	5.5	5.8	8.7	8.4	—

D. Summary by Textbook used (Eells); medians.

NAME OF TEXTBOOK	NUMBER OF SCHOOLS	NUMBER OF STUDENTS	ADDITION AND SUBTRACTION SERIES A 12 WEEKS	EQUATION AND FORMULA SERIES A 12 WEEKS	MULTIPLICATION AND DIVISION SERIES A 24 WEEKS	PROBLEMS SERIES B 24 WEEKS	GRAPHS SERIES A 24 WEEKS
Wells & Hart	26	1917	5.7	5.8	7.4	7.4	5.5
Hawkes, Luby, & Touton	18	1193	5.5	5.5	7.3	6.7	4.7
Slaught- Lennes	4	131	4.9	5.1	5.7	5.8	3.3
Edgerton- Carpenter	3	348	4.2	6.0	6.1	6.9	—
Durrell- Arnold	3	96	5.1	5.2	6.5	7.3	5.3
Sykes-Com- stock	2	72	4.5	5.7	5.5	7.3	6.1
Stone-Mills	2	43	5.1	4.9	—	—	—
Milne	2	26	4.9	5.9	7.0	6.2	—

Douglass Standard Diagnostic Tests for Elementary Algebra

Description of the tests. Series A is designed to measure the four fundamental operations of elementary algebra. Each 4-page folder contains four tests as follows:

- Test I. Collection of Terms
- Test II. Multiplication
- Test III. Division
- Test IV. Solution of Simple Equations

Ten exercises are given in each test in order of increasing difficulty, covering the chief phases of each operation in algebra. The four operations chosen for Series A are based on the suggestions of a representative group of fifty prominent teachers of secondary mathematics. They conform also to the recommendations of the National Committee on Mathematical Requirements regarding courses of study. Series A should be given near the end of the first semester. Near the end of the first year of algebra Series B should be given. It consists of seven tests, as follows:

- Test I. Fractions
- Test II. Factoring
- Test III. Fractional Equations
- Test IV. Simultaneous Equations
- Test V. Graphs
- Test VI. Square Root, Radicals, and Exponents
- Test VII. Quadratic Equations

Administration and scoring. Detailed instructions for administering these tests accompany the tests. Series A requires 34 minutes' working time, and Series B 102 minutes. In scoring, a problem is either right or wrong in conformance with rules given in the directions. Subjective judgment does not enter.

Interpretation and utilization of results. These tests may be considered diagnostic of individual performance only when a complete series is given; detailed use of the test by parts should be confined to class diagnosis. However,

the part-scores, together with a tabulation of specific errors, will indicate the skills in need of remedial teaching. A study by M. L. Fossler¹ indicates the type of errors most frequently found by means of the Douglass tests, Series A. Two hundred high school pupils in three high schools who had just finished first-year algebra made 2254 errors, as follows:

TYPE OF ERROR	NUMBER OF ERRORS	PER CENT OF TOTAL ERRORS
Exponents	606	26.9
Signs	550	24.4
Operation of Problem	521	23.1
Coefficients	218	9.7
Terms	134	5.9
Miscellaneous	128	5.7
Letters	97	4.3

Validity and reliability. Standardization of Series A was based on about 1000 pupils. Data on reliability are given in Table 7.

TABLE 7

DATA ON RELIABILITY OF THE DOUGLASS STANDARD DIAGNOSTIC TESTS FOR ELEMENTARY ALGEBRA

r	N	S.D.	P.E. score	P.E. $_{\infty,1}$	P.E. score S.D.	P.E. $_{\infty,1}$ S.D.	NATURE OF GROUP
.80	175	5.20	1.6	1.4	.31	.27	First-Year Class
.84	43	4.89	1.3	1.2	.27	.25	First-Year Class

¹ Fossler, M. L., *A Study of the Errors Made by Students Who Have Completed First-Year Algebra as Shown by Douglass Standard Diagnostic Tests for Elementary Algebra*. M. A. Thesis. (University of Iowa, Iowa City, 1924; 24 pages.)

Norms. Table 8 gives the tentative norms for the Douglass tests.

TABLE 8

TEST NUMBER	I	II	III	IV	V	VI	VII
Series A	7.8	7.1	6.5	7.3	—	—	—
Series B	2.4	4.1	3.1	3.6	2.5	2.7	3.4

OTHER TESTS IN ALGEBRA

The Illinois Standardized Algebra Tests devised by W. S. Monroe and L. W. Williams are unique. All the exercises in these tests consist of simple equations of four general types. Monroe represents the types as follows:

$$\text{Test I. } \pm ax \pm bx = \pm c$$

$$\text{Test II. } \pm ax \pm c = \pm bx \pm d$$

$$\text{Test III. } \pm k(\pm ax \pm c) = \pm bx \pm d$$

$$\text{Test IV. } \pm \frac{\pm ax \pm c}{n} = \pm \frac{\pm bx \pm d}{m}$$

Exercises are arranged on the cycle principle; i.e., those of the same type in each test recur at regular intervals, and each type is met the same number of times. This whole test rests on the assumption that achievement in the fundamentals of algebra may be measured by varying the sign combinations in linear equations. The reliability of the Illinois Tests is given in Table 9.

TABLE 9

DATA ON THE RELIABILITY OF THE ILLINOIS STANDARDIZED ALGEBRA TESTS

r	N	S.D.	P.E. score	P.E. $_{\infty,1}$	P.E. score S.D.	P.E. $_{\infty,1}$ S.D.	NATURE OF GROUP
.88	38	9.32	.22	2.0	.23	.22	Grade IX pupils

Thurstone Vocational Guidance Tests — Algebra

The Algebra Test of the Thurstone Vocational Guidance Tests¹ requires 30 minutes' working time. It is designed for students about to enter an engineering course in college, and is standardized on the basis of the performance of 7000 college freshmen. It is of prognostic value with respect to probable success in technical work, the correlation between test scores and average freshman scholarship being .42.

Algebra Test Intercorrelations

Table 10 gives (A) intercorrelations among algebra tests, and (B) correlations between algebra and other factors. These values indicate the extent to which the different tests measure the same functions.

TABLE 10

(A)	ALGEBRA TEST INTERCORRELATIONS	r	N
Hotz Algebra (Series A) vs. Douglass Algebra	.72	175	
Hotz Algebra (Series B) vs. Douglass Algebra	.65	38	
Hotz Algebra (Series A) vs. Illinois Algebra	.51	38	
Hotz Algebra (Series A) vs. Teachers' Estimates	.40	38	
Douglass Algebra (Form I) vs. Illinois Algebra	.52	38	
Douglass Algebra (Form I) vs. Teachers' Estimates	.64	38	
Illinois Algebra vs. Teachers' Estimates	.45	38	

(Table continued on next page)

¹ See Chapter X for a description of these tests.

(B) CORRELATIONS BETWEEN ALGEBRA AND OTHER FACTORS (OBTAINED WITH 175 STUDENTS IN THREE HIGH SCHOOLS) ¹		
FACTOR	HOW MEASURED	CORRELATION WITH ALGEBRA
Algebra	Hotz + Douglass	.94 (reliability)
Arithmetic	Stanford Arithmetic Examination	.59
Reading	Stanford Reading Examination	.32
Intelligence	Terman Group + Otis Advanced	.48
Chronological Age	Expressed in months	.12
Teachers' Ratings	Rating Scale for Interest and Persistence	.40

Iowa Placement Examinations²

Two mathematics examinations are available in the Iowa Placement series: Mathematics Aptitude, Revised (2 forms), and Mathematics Training, Revised (2 forms). These Examinations comprise the following material:

Mathematics Aptitude, MA-1, Revised

Part 1 consists of 15 arithmetic and algebraic number series similar to those commonly found in intelligence tests. Part 2 (15 items) is a test of constructive imagination. Solution of the problems depends upon the pupil's ability to visualize geometric figures and to see the relations involved. Part 3 (20 items) is a test of logic. This test meets the criteria usually set up for pure intelligence tests and is included in Mathematics Aptitude because success in mathematics is in part a function of one's ability to deal with highly abstract material. Part 4 (15 items) is a test of mathematical read-

¹ McCoy, J. P., *An Analysis of Algebraic Abilities*. Ph.D. Thesis. (University of Iowa, Iowa City, 1924.)

² See Chapter XI for a complete description of the validity, reliability, and utility of these examinations.

ing comprehension. The material is drawn from a standard textbook in calculus and is fairly representative of the type of material which the student in mathematics will encounter. The form is modeled after the Iowa Comprehension Test. The total working time of MA-1, Revised, is 40 minutes.

Mathematics Training, MT-1, Revised

In Part 1, 20 problems are devoted to the fundamentals of arithmetic, each bringing out a different skill. They are drawn for the most part from teachers' experience. Part 2 is a sample of 20 problems in formal algebra. All the items included recur constantly in algebraic work. Part 3 (40 items) measures the fundamentals of geometry. This part is of the true-false type, but it involves knowledge of geometric relationships in that the student is in many cases forced to draw a figure in order to give a correct response. Part 4 consists of 15 algebraic reasoning problems in which the mechanical computation is reduced to a minimum. The total working time of MT-1, Revised, is 40 minutes.

Remedial procedures in algebra. The particular applications of algebra tests have previously been indicated. With respect to the formal aspect of algebra, the tests described can be depended upon for valuable assistance in measuring the class achievement and locating pupil weaknesses. To some degree the teacher can now locate teacher weakness and textbook weakness. For example, adequate knowledge of what constitutes a real difficulty for the pupil should lead to corresponding improvement in teaching methods. Tests should be given regularly, and class and pupil records kept in accessible form. The teacher can often enlist the active interest of the pupils in their performance.

There are additional purposes which an algebra test can be made to serve. These are summed up in the 1923 Report

of the National Committee on Mathematical Requirements (page 362) :

Tests are enabling teachers to know such things as the following :

- (1) That certain parts of algebra are much more difficult than other parts.
- (2) That a pupil's difficulty with a given topic is often due to neglect on the part of the teacher to provide proper instruction in certain small details of that topic.
- (3) That certain topics, traditionally thought to be difficult, are not so difficult, and in fact are comparatively easy if properly taught.
- (4) That teachers throughout the country have not been at all uniform in the amount of time or emphasis they have given to the various topics of first-year algebra.
- (5) That speed and accuracy have certain close relationships.

Osburn¹ emphasizes the necessity for (1) definite identification of the difficulty encountered by the pupil, and (2) abundant drill exercises to take care of them. Ten factors which must be taken into account to avoid failures in algebra, but which teachers are likely to overlook, are the following : (1) vocabulary knowledge; (2) comprehension of symbols; (3) utilization of all necessary problem-data; (4) use of inverse relations; (5) prohibition of harmful transfer; (6) horizontal addition and subtraction; (7) reading between the lines; (8) help in contradictory situations; (9) generalization; (10) proportion. These considerations enter to some extent in all branches of mathematics.

II. GEOMETRY

Minnick Geometry Tests

Description of the tests. Four tests, A, B, C, and D, are provided, each printed in a separate folder. Test A consists of five stated propositions for each of which the pupil

¹ Osburn, W. J., "Ten Reasons Why Pupils Fail in Mathematics." *The Mathematics Teacher*, Vol. XVIII, No. 4 (April, 1925), pages 234-238.

is to construct the figure. In Test B the figure is drawn and the theorem is stated. The pupil is required to state what is given and what is to be proved. In Test C a figure is given, certain facts about it are stated, and the pupil is requested to give as many more facts about the figure as he can. In Test D figures are drawn for a series of exercises, facts are given, and the hypotheses stated. The pupil is asked to supply the proof. Only one form of each test is available.

What the tests measure. The tests do not attempt to go beyond the formal phase of geometry. Minnick divides formal geometry into three chief divisions: the demonstration of theorems, the construction of figures under given conditions, and the solution of numerical problems. Certain abilities involved in the demonstration of theorems are covered by the tests. They follow the order common to geometrical demonstrations. The usual steps are:

- (1) Drawing the figure described in the theorem;
- (2) Stating the hypothesis and the conclusion;
- (3) The mustering of other known, relevant facts;
- (4) The proper selection and arrangement of these facts; i.e., pointing them to the conclusion.

The four tests, A, B, C, and D, parallel respectively these steps. The sampling, however, is limited even for the particular scope outlined.

Administration and scoring. No difficulty will be encountered in administering the test to a group of pupils. The test can be given any time after the completion of the first two books in plane geometry. The work is done on the test blanks, each pupil being provided with a pencil. Each test requires 30 minutes' working time. Before beginning work pupils are allowed to ask questions concerning the instructions. A special method of scoring characterizes the

Minnick tests. Two scores are kept. One is based on the number of things done correctly and is called the positive score. The other indicates the number of incorrect or unnecessary statements and is called the negative score. The latter adds to the diagnostic value of the tests. It gives the teacher information on the types of errors introduced and on the special difficulties of particular pupils. Standards are provided for both scores. The dual aspect of the scoring system and the number of possible correct solutions for each exercise make scoring a tedious business. There are listed on the Class Record Sheet seven general directions and twelve special directions to be followed in scoring the complete series. Specific values are assigned to each correct operation in each exercise. The teacher should familiarize himself thoroughly with the scoring method in order to avoid as far as possible the effects of personal opinion.

Interpretation and utilization of results. These tests were designed to diagnose pupil difficulties in geometry. Attainment of the standard scores will be conditioned largely by the importance attached to the formal aspects of geometry. The complexity of the scoring system introduces errors which lessen the value of the norms. In some cases the directions to students are not very explicit and the resulting variety of performances is differently interpreted by different scorers. At Plattsburg, New York, three judges differed by as much as 22 points in the positive scores and 11 points in the negative scores.¹

Norms. The Class Record Sheets provide space for both positive and negative scores. Standards (medians, 25-percentiles, and 75-percentiles) are furnished for both types of scores. But such norms should not be taken too literally.

¹ Morrison, J. C., "The Use of Standard Tests and Scales in the Plattsburg High School." *University of the State of New York Bulletin*, No. 784 (July 15, 1923); 45 pages

They were obtained from sixty-three schools. Table 11 gives an idea of the underlying situation.

TABLE 11
NORMS FOR THE MINNICK GEOMETRY TESTS

TEST	POSITIVE SCORES			TEST	NEGATIVE SCORES		
	Lowest Median	Highest Median	"Standard Score"		Lowest Median	Highest Median	"Standard Score"
A	50.5	78.7	62.5	A	11.8	4.8	7.1
B	38.5	80.9	69.3	B	4.5	2.0	3.5
C	29.0	67.0	50.6	C	7.3	2.4	4.1
D	54.7	80.5	73.3	D	3.7	1.5	2.6

The Illinois grade norms given agree rather well for Tests A and B; but the Illinois medians for Tests C and D (Grade X) are close to Minnick's 75-percentile scores.

TABLE 12
GRADE NORMS FOR MINNICK GEOMETRY TESTS¹

	POSITIVE SCORES		NEGATIVE SCORES	
	Grade		Grade	
	X	XI	X	XI
Test A				
Number of pupils	126	66	126	60
25-percentile	53.3	43.8	2.4	1.1
Median	63.0	58.0	4.1	2.6
75-percentile	67.2	69.2	6.6	5.4

(Table continued on next page)

¹ Monroe, W. S., "Report of Division of Educational Tests for 1919-1920." *University of Illinois Bulletin*, Vol. XVIII, No. 21 (January 24, 1921); 64 pages.

TABLE 12 (Continued)

Test B				
Number of pupils	167	66	167	66
25-percentile	55.2	55.5	1.1	1.0
Median	69.6	67.1	2.3	2.0
75-percentile	81.3	83.6	3.9	3.9
Test C				
Number of pupils	154	65	154	63
25-percentile	52.2	55.6	1.9	1.4
Median	64.1	64.7	3.8	3.9
75-percentile	77.2	77.9	7.1	5.7
Test D				
Number of pupils	155	68	155	54
25-percentile	68.0	75.0	.8	.8
Median	85.5	89.2	1.6	1.6
75-percentile	92.9	98.3	2.3	3.2

Validity and reliability. The test was standardized with about 1000 children in 63 schools. Weights attached to the different operations were based on the average per cent of correct statements given by the pupils. The pupils had completed the first two books of geometry. Data on the reliability of the Minnick Geometry Tests are given in Table 13.

TABLE 13

DATA ON THE RELIABILITY OF THE MINNICK GEOMETRY TESTS

A	$N = 61$			
	$r_{AB} = .71$	$r_{AC} = .51$	$r_{AD} = .55$	
B	$r_{BC} = .63$	$r_{BD} = .60$	$r_{CD} = .80$	
B	r	N	S.D.	P.E. score
	.63	61	19.6	8.0
				P.E. $\infty, 1$
				.64
				P.E. score / S.D.
				.41
				P.E. $\infty, 1$ / S.D.
				.32

Schorling-Sanford Achievement Test in Plane Geometry

Description of the test. Two equivalent forms, A and B, are available, each consisting of five parts of twelve questions each, organized about the following topics:

- Part I. Sentence Completion (Factual Material)
- Part II. Drawing Conclusions from Given Data
- Part III. Judging the Correctness of Conclusions
- Part IV. Analyzing Constructions
- Part V. Computations (Angles, Areas, etc.)

Geometrical figures are printed on the page for Parts II to V, and a fore-exercise is provided for all parts.

What the test measures. This test is designed only as a final achievement examination in plane geometry. The subject matter of the test is drawn from all five books of geometry. It is not intended for pupil diagnosis throughout the year.

Administration and scoring. Each form requires 52 minutes' working time. A definite time allowance is made for each part, and all pupils must work on the same part at the same time. Since the fore-exercises require about 10 minutes, the teacher should allow two class-hours for the testing. Each test-part yields a maximum score of 12 points, one point for each question answered correctly except for Part III, which is a 3-response type corrected for chance. The scoring is fairly objective and proceeds rapidly.

Interpretation and utilization of results. The Schorling-Sanford test can be used (1) to make comparisons of the work of various classes and schools, and (2) in the improvement of teaching technique. Tabulation of the errors made will enable the teacher to revise his instructional units, drill material, etc., in order to bring the work of the course in better harmony with the principles of learning.

Validity and reliability. The items were originally selected by Raleigh Schorling in 1921. They have subsequently been selected in accordance with the recommendations of the National Committee on Mathematical Requirements.

Reliability coefficients (Form A *vs.* Form B) range from .46 to .88 for single classes; the reliability for 284 cases (in 12 high school classes, by class) is .72.

Norms. Tentative percentile scores and frequency distributions are given in the Manual of Directions. The number of cases is 695 for Form A and 290 for Form B. More detailed standards are necessary in order to judge accurately class performance.

OTHER TESTS IN GEOMETRY¹

The Columbia Research Bureau Plane Geometry Test is designed as a final examination in geometry for use in high schools and as a college entrance test. It consists of two parts, a 20-minute true-false test of 65 items and a 40-minute problem test of 30 items. The Manual is accompanied by a supplement containing directions for the testing of loci, converses, definitions, and ability in demonstration.

The Geometry Test of the Thurstone Vocational Guidance Tests² requires 30 minutes' working time. It is designed for beginning college students and emphasizes geometrical construction. It yields a correlation of .30 with first-year engineering scholarship. Complete directions for administering, scoring, and interpreting this test are given in the Manual of Directions.

Remedial procedures in geometry. The teacher should confine his attention to correcting the pupil faults discovered not only by standard tests but by objective examinations which he may devise. The procedure here is the same as for

¹ See Chapter XI for the Columbia Research Bureau Plane Geometry Test.

² See Chapter X for a complete description of these tests.

algebra. The tests now available in geometry are too unreliable to furnish more than a general estimate of class and school achievement. At the present time the Report of the National Committee on Mathematical Requirements offers the best guide in the selection of content material for tests and examinations. Remedial measures should be undertaken in order to fill gaps in essential material, and above all to insure that pupils in geometry really get something from the work. Thus it is possible to turn geometry into an effective and interesting laboratory science. Austin's plan¹ "proposes to introduce the pupil first to concrete form. A completed drawing some one else has made is not even shown him. . . . He performs an experiment in which a law is so evident that its operation cannot escape his attention. Then follows the ordinary proof of the fact observed to establish it as a general truth." This procedure emphasizes independent thinking. It is doubtful if many of the small skills in plane geometry which involve memorized factual material are worth painstaking efforts to correct them. Remedial measures should be devoted to stimulation of independent, critical attack upon the problems of geometry; to pupil growth, rather than to knowledge accumulation.

III. OTHER MATHEMATICS TESTS

The Rogers Test of Mathematical Ability, sometimes called the "Rogers Sextet," is designed to measure in advance a pupil's capacity to do successful work in high school mathematics. For the most part the material covered is new to the pupils taking the test, thus putting the prognosis essentially upon inherited capacities and thoroughly learned problem-solving habits. The total working time is $1\frac{1}{2}$ hours.

¹ Austin, W. A., "Geometry a Laboratory Science." *School Science and Mathematics*, Vol. XXIV (January, 1924), pages 58-71.

It may be given at the end of junior high school or upon completion of one year of high school.

The Rogers tests have been used chiefly to advise pupils concerning further study in mathematics, to furnish a check on school marks in mathematics, and to section classes on the basis of mathematical ability. Wherever possible they should be used in conjunction with a standard intelligence test and school marks. Norms for various types of schools are given in the Manual of Directions. Data on reliability are given by Agnes L. Rogers in *Teachers College, Columbia University, Contributions to Education*, No. 89 (1918). On 28 pupils in Grade IX of the University of Iowa High School the writers obtained the reliability data in Table 14.

TABLE 14

DATA ON THE RELIABILITY OF THE ROGERS TEST OF
MATHEMATICAL ABILITY

r	N	S.D.	P.E. score	P.E. $_{\infty.1}$	P.E. score S.D.	P.E. $_{\infty.1}$ S.D.	NATURE OF GROUP
.82	28	34.15	9.8	8.8	.29	.26	Grade IX, Uni- versity of Iowa High School

The Kelley Mathematical Values Test was devised by T. L. Kelley at Columbia University, in 1917, in his attempt to answer the question, "What are the values of high school algebra, and how are they to be measured?" By means of this test it is possible to estimate the broader interests and applications developed by high school mathematics courses.

Test Materials

Hotz Algebra Scales. By HENRY G. HOTZ. For pupil: One copy of each scale selected, each scale 70 cents per 100, except Graph Scale, which is \$1.25 per 100. For examiner: One copy of Manual of Directions for First-Year Algebra Scales, 75 cents. Bureau of Publications, Teachers College, Columbia University, New York. Published also by Public School Publishing Company, Bloomington, Illinois.

Douglass Standard Diagnostic Tests for Elementary Algebra. By H. R. DOUGLASS. Series A: Tests I, II, III, IV (Form I or II), \$1.60 per 100; Series B, \$3.50 per 100, including Key and Class Record Sheet. Published by University of Oregon, Eugene, Oregon.

Illinois Standardized Algebra Tests. By W. S. MONROE and L. W. WILLIAMS. \$2.50 per 100. Specimen set, 15 cents. Public School Publishing Company, Bloomington, Illinois.

Minnick Geometry Tests. By J. H. MINNICK. Tests A, B, C, and D, each \$2.50 per 100. Specimen set, 20 cents. Public School Publishing Company, Bloomington, Illinois.

Schorling-Sanford Test in Plane Geometry. By RALEIGH SCHORLING and VERA SANFORD. For examiner: One Manual of Directions and set of stencils, 50 cents. For pupil: Test Booklet, \$7.00 per 100. Specimen set, 10 cents. Bureau of Publications, Teachers College, Columbia University, New York.

Hawkes-Wood Plane Geometry Examination. By H. E. HAWKES and BEN D. WOOD, Columbia University, New York. An experimental edition; write authors of the test.

Thurstone Vocational Guidance Tests. By L. L. THURSTONE. Algebra Test, \$1.00 per package of 25, with Key and Record Sheet; Geometry Test, \$1.00 per package of 25, with Record Sheet (no Key required). Manual of Directions, 20 cents. World Book Company, Yonkers-on-Hudson, New York.

Rogers Tests for Diagnosing Mathematical Ability. By AGNES L. ROGERS. For examiner: One Manual of Directions and set of stencils, 50 cents. For pupil: Test Booklet, \$7.00 per 100. Specimen set, 10 cents. Bureau of Publications, Teachers College, Columbia University, New York.

Kelley Mathematical Values Test. By T. L. KELLEY. For examiner: One set of scales, 40 cents; one copy *Teachers College Record*, May, 1920, 40 cents. For pupil: Test Blank, 5 cents. Bureau of Publications, Teachers College, Columbia University, New York.

Iowa Placement Examinations.

Mathematics Aptitude, MA-1, Revised, Forms A and B

Mathematics Training, MT-1, Revised, Forms A and B

Each form \$3.50 per 100, with Manual and Key. Bureau of Educational Research and Service, University of Iowa, Iowa City, Iowa.

References

AUSTIN, W. A. "Geometry a Laboratory Science." *School Science and Mathematics*, Vol. XXIV (January, 1924), pages 58-71.

BARBER, H. C. "Real Improvement in Algebra Teaching." *The Mathematics Teacher*, Vol. XVIII (October, 1925), pages 364-374.

DALMAN, M. A. "Hurdles. A Series of Calibrated Objective Tests in First-Year Algebra." *Journal of Educational Research*, Vol. I (January, 1920), pages 47-62.

DEAM, THOMAS M. "Diagnostic Algebra Tests and Remedial Measures." *School Review*, Vol. XXXI (1923), pages 376-379.

DOUGLASS, H. R. "The Derivation and Standardization of a Series of Diagnostic Tests for the Fundamentals of Elementary Algebra." *University of Oregon Publication*, Vol. I, No. 8 (April, 1921); 48 pages. University of Oregon, Eugene, Oregon.

— "A Series of Standardized Diagnostic Tests for the Fundamentals of Elementary Algebra." *Journal of Educational Research*, Vol. IV (December, 1921), pages 396-403.

EELLS, W. C. "What Amount of Algebra Is Retained by College Freshmen?" *The Mathematics Teacher*, Vol. XVIII, No. 4 (April, 1925), pages 219-225.

— "Hotz Algebra Scales in the Pacific Northwest." *The Mathematics Teacher*, Vol. XVIII (November, 1925), pages 418-427.

HARRIS, ELEANORA, and BREED, F. A. "Comparative Validity of the Hotz Scales and the Rugg-Clark Tests in Algebra." *Journal of Educational Research*, Vol. VI (December, 1922), pages 393-411.

HOTZ, H. G. "First-Year Algebra Scales." *Teachers College Contributions to Education*, No. 90 (1918); 87 pages. Columbia University, New York.

— *Teacher's Manual for First-Year Algebra Scales*. Bureau of Publications, Teachers College, Columbia University, New York; 1922.

INGLIS, A. *Principles of Secondary Education*, Chapter XIV. Houghton Mifflin Company, Boston; 1918.

KELLEY, T. L. "Values in High School Algebra and Their Measurement." *Teachers College Record*, Vol. XXI, No. 3 (May, 1920), pages 246-290.

MENSENKAMP, L. E. "Tests of Mathematical Ability and Their Prognostic Values; A Discussion of the Rogers Tests." *School Science and Mathematics*, Vol. XXI (February, 1921), pages 150-162.

MINNICK, J. H. *An Investigation of Certain Abilities Fundamental to the Study of Geometry*. University of Pennsylvania, Philadelphia, Pennsylvania; 1918.

— "Certain Abilities Fundamental to the Study of Geometry." *Journal of Educational Psychology*, Vol. IX (February, 1918), pages 83-90.

MONROE, W. S. "Report of Division of Educational Tests for 1919-1920." *University of Illinois Bulletin*, Vol. XVIII, No. 21 (1921); 64 pages. University of Illinois, Urbana, Illinois.

OSBURN, W. J. "Ten Reasons Why Pupils Fail in Mathematics." *The Mathematics Teacher*, Vol. XVIII, No. 4 (April, 1925), pages 234-238.

PERRY, W. M. "Student Difficulties in Exercises in Geometry." *The Mathematics Teacher*, Vol. XVIII (February, 1925), pages 79-82.

PIERCE, P. R. "Report of an Experiment in Correlated Mathematics in a Large High School." *School Science and Mathematics*, Vol. XXV (October, 1925), pages 681-684.

Problem of Mathematics in Secondary Education. Report of the Commission on the Reorganization of Secondary Education, Appointed by the National Education Association. Bulletin No. 1 (1920). Department of the Interior, Bureau of Education, Washington, D. C.

Reorganization of Mathematics in Secondary Education. A Summary of the Report by the National Committee on Mathematical Requirements. Bulletin No. 32 (1921). Department of the Interior, Bureau of Education, Washington, D. C.

ROGERS, AGNES L. *Experimental Tests of Mathematical Ability and Their Prognostic Value.* Bureau of Publications, Teachers College, Columbia University, New York; 1918.

— "Directions for Using the Rogers Test of Mathematical Ability." Bureau of Publications, Teachers College, Columbia University, New York; 1921.

— "Tests of Mathematical Ability — Their Scope and Significance." *The Mathematics Teacher*, Vol. XI (June, 1919), pages 145-163.

RUGG, H. O., and CLARK, JOHN R. "Standardized Tests and the Improvement of Teaching in First-Year Algebra." *School Review*, Vol. XXV (February, 1917), pages 113-132; (March, 1917), pages 196-213.

— "A Coöperative Investigation in the Testing and Experimental Teaching of First-Year Algebra." *School Review*, Vol. XXV (May, 1917), pages 346-349.

— "The Improvement of Ability in the Use of the Formal Operations of Algebra by Means of Formal Practice Exercises." *School Review*, Vol. XXV (October, 1917), pages 546-554.

SANFORD, VERA A. "A New-Type Final Geometry Examination." *The Mathematics Teacher*, Vol. XVIII (January, 1925), pages 22-36.

SCHORLING, RALEIGH, and CLARK, JOHN R. "A Program of Investigation and Coöperative Experimentation in the Mathematics of the Seventh, Eighth, and Ninth School Years." *The Mathematics Teacher*, Vol. XIV (May, 1921), pages 264-275.

SCHREIBER, EDWIN W. "A Study of the Factors of Success in First-Year Algebra." *The Mathematics Teacher*, Vol. XVIII (February, 1925), pages 65-78; (March, 1925), pages 141-163.

STODDARD, GEORGE D. "Iowa Placement Examinations." *University of Iowa Studies in Education*, Vol. III, No. 2 (August 15, 1925); 103 pages.

SYMONDS, P. M. "The Psychology of Errors in Algebra." *The Mathematics Teacher*, Vol. XV (February, 1922), pages 93-104.

THORNDIKE, E. L. "The Nature of Algebraic Abilities." *The Mathematics Teacher*, Vol. XV (January, 1922), pages 6-15; (February, 1922), pages 79-92.

— and WOODYARD, ELLA. "The Uses of Algebra in Study and Reading." *School Science and Mathematics*, Vol. XXII (May, 1922), pages 405-415; (June, 1922), pages 514-522.

WALKER, HELEN M. "What the Tests Do Not Test." *The Mathematics Teacher*, Vol. XVIII (January, 1925), pages 45-53.

WILLIAMS, L. W. "Illinois Standardized Algebra Test." *Journal of Educational Research*, Vol. III (January, 1921), pages 75-76.

YOUNG, J. W., et al. *The Reorganization of Mathematics in Secondary Education*. A Report by the National Committee on Mathematical Requirements of the Mathematical Association of America, Inc. (1923). J. W. Young, Dartmouth College, Hanover, New Hampshire, Chairman of the Committee.

CHAPTER SIX

ENGLISH (LANGUAGE, GRAMMAR, SPELLING, READING, AND COMPOSITION)

I. LANGUAGE AND GRAMMAR TESTS

Introduction. Standard tests in language and grammar for the high school range are limited in number, and are given over to measurement of only the most formal aspects of the subject. The chief limitations of these tests may be summarized as follows:

1. They fail to measure adequately the functional side of language usage;
2. They offer, for the most part, very inadequate samplings of the field of language and grammar;
3. They are too unreliable (except for grammar principles) for individual measurement and diagnosis;
4. The norms available fail to indicate the progress of pupil or class; according to the tests available, retrogression is not infrequent;
5. The exact experimental and statistical investigations by which the tests were validated are not made available to the general reader or test user.

Some of these deficiencies are characteristic of the majority of standard tests, but it appears that English testing presents special difficulties. The National Joint Committee on English¹ has pointed out that "all expression in writing demands correctness as to formal details; namely, a legible and firm handwriting, correct spelling, correctness in grammar

¹ Hoscic, J. F. (Chairman), *Reorganization of English in Secondary Schools*. Report by the National Joint Committee on English, Representing the Commission on the Reorganization of Secondary Education of the National Education Association and the National Council on English. Bulletin No. 2 (1917), page 31 (Department of the Interior, Bureau of Education, Washington, D. C.).

and idiom, and observance of the ordinary rules for capitals and marks of punctuation; [and that] the writer should make an effort to gain an enlarged vocabulary, a concise and vigorous style, and firmness and flexibility in constructing sentences and paragraphs." The same report¹ indicates the place of grammar in the English program: "A sane attitude toward the teaching of grammar would seem to be to find out what parts and aspects of the subject have actual value to children in enabling them to improve their speaking, writing, and reading, to teach these parts according to modern scientific methods, and to ignore any and all portions of the conventional school grammar that fall outside these categories."

The tests now available will give assistance in the improvement of teaching technique; if they are weighted toward the formal aspects of English, it is because testing there is easier, more reliable, and more readily adjusted to teaching aims.

Kirby Grammar Test

Description of the test. The Kirby Grammar Test is designed to measure two specific abilities in English:

- (1) The ability to choose correct English usage, and
- (2) The ability to recognize the grammatical principle by virtue of which the usage is considered correct.

There are five sections to the test, the first two dealing with pronouns, the second two with verbs, and the fifth with miscellaneous items. The sentences are arranged in a left-hand column, while the principles are to the right, in scrambled order. In effect, the pupil is confronted with a 2-response situation in the language part, and a matching-test in locating the correct grammatical principle. The examples in the fore-exercise make this clear:

¹ *Op. cit.*, page 37.

<i>Sentences</i>	<i>Principles</i>
<input type="checkbox"/> 1. (Whom) (Who) did you meet?	a. The indirect object is in the objective case.
<input type="checkbox"/> 2. He told John and (I) (me) an interesting story.	b. The subject of a verb is in the nominative case. c. The object of a verb is in the objective case.

The student is required to cross out the word in parentheses that is incorrect, and to indicate in the square the letter opposite the principle which applies in each case.

What the test measures. Kirby secured his material from a large number of actual mistakes made by pupils. The language errors included are the most common ones. The rules or principles are those commonly fitted to the language situations. No extra rules are given, and the pupil should have no difficulty in checking the proper rule, if he knows it in a general way. The list is not at all exhaustive, but it has this advantage: it can be given in the ordinary class period. The test is designed for Grades VII to XII.

Administration and scoring. The teacher will have no difficulty in administering and scoring the Kirby Grammar Test. Each form requires 35 minutes' working time. Scoring is completely objective. Norms include the number attempted and the number right.

Interpretation and utilization of results. Teachers of English are frequently disturbed by the fact that some pupils in the higher grades show an amazing ignorance of fundamental usages. The Kirby test is a convenient device to discover what weaknesses remain. It helps in determining where the emphasis is now needed and how much time should still be devoted to grammar. Its utility here is essentially diagnostic. In a very real sense the high school period is a "last chance" to iron out these deep-lying defects in grammar. What the class as a whole does is of secondary import-

tance, for the major objectives in the teaching of English are now expressed in the intricacies of good English composition. To the questions, How much grammar? and What grammar? Kirby makes this tentative answer: Enough grammar to "reason himself out of errors," and such grammar that principles are made intelligible. Having located pupil weaknesses by means of the Kirby Grammar Test and additional objective examinations designed to cover the whole field of English, the teacher must attempt to build up in his pupils correct *habits* in grammatical usage.

Validity and reliability. An exhaustive study would be needed to validate properly an adequate language-grammar test. The teacher will run across many errors not touched upon by this test, but those covered by the test are the most common and apparently the most difficult to eradicate. Data for the reliability of the Kirby Grammar Test are given in Table 15. It should be noted that the reliability for principles is high, but is not satisfactory for sentences when individual pupils are to be dealt with.

TABLE 15

DATA ON THE RELIABILITY OF THE KIRBY GRAMMAR TEST

<i>r</i>	<i>N</i>	S.D.	P.E. score	P.E. _{∞.1}	P.E. score S.D.	P.E. _{∞.1} S.D.	NATURE OF GROUP
<i>Principles</i>							
.87	30	7.7	1.9	1.7	.24	.23	Grade VII
.90	18	5.0	1.1	1.0	.21	.20	Grade VIII
.91	128	9.1	1.9	1.8	.20	.19	Grades VII-XII
<i>Sentences</i>							
.60	32	4.0	1.7	1.3	.42	.33	Grade VII
.43	18	3.4	1.7	1.1	.50	.33	Grade VIII
.70	136	4.3	1.6	1.3	.37	.31	Grades VII-XII

Norms. The medians for number of principles correct and sentences correct are given in the Manual of Directions for Grades VII to XII. There is so little change from grade to grade in the high school range that this test does not afford a satisfactory measure of class progress.

OTHER LANGUAGE-GRAMMAR TESTS

Wilson Language Error Test

Description of the test. The Wilson Language Error Test is of special value in Grades III to VIII, but can also be used in high school. Its purpose is "to discover the pupil's ability to recognize and avoid the common language errors." Three stories are printed in one booklet, each containing 28 errors to be corrected by the pupil. The stories are of equal difficulty, and may be considered as three forms of the same test. A second set of stories is not available. Progress may be measured directly by giving the three stories at the beginning, the middle, and the end of the school year. The procedure parallels that followed by the pupil in correcting his own themes. Corrections are indicated by crossing out the word or phrase containing the error, and inserting above it the correct form. Scoring is quickly done. In high school classes it should be feasible to have the students themselves score the tests, preferably after exchanging papers with classmates. For high school pupils the working time of the test is from 5 to 10 minutes.

Utilization of results. In the Wilson Language Error Test the emphasis is placed directly on language rather than on technical knowledge of grammar. The norms for high school grades do not show much spread. These medians are, respectively, 23, 24, 25, and 26; and the probable error of a score is given as 2.2 score points. The true value of this test lies in its diagnostic utility, and this is rather limited

because of the brevity of the test. Additional data on the reliability of the Wilson Language Error Test are given in Table 16.

TABLE 16

DATA ON THE RELIABILITY OF THE WILSON LANGUAGE ERROR TEST

r story 1 vs. story 2 .80
 r story 1 vs. story 3 .73
 r story 2 vs. story 3 .66 *N* 38 (Grade IX, H. S.)

r (Av.)	<i>N</i>	S.D. (Av.)	P.E. score	P.E. ∞ .1	P.E.-score S.D.	P.E. ∞ .1 S.D.	NATURE OF GROUP
.73	38	4.58	1.6	1.4	.35	.30	Grade IX, H. S.
.90	103	7.10	1.5	1.4	.21	.20	Grades III-VIII

Cross English Test

Description of the test. The Cross English Test is published in three equivalent forms: A, B, and C. Each form contains the following material:

Part I.	Spelling (32 items)
Part II.	Pronunciation (32 items)
Part III.	Recognizing a Sentence (40 items)
Part IV.	Punctuation (15 items)
Part V.	Verb Forms (16 items)
Part VI.	Pronoun Forms (12 items)
Part VII.	Idiomatic Expressions (10 items)
Part VIII.	Miscellaneous Faulty Expressions (15 items)

The test is designed chiefly for high school seniors and college freshmen, but can be given throughout high school. Its working time is 45 minutes.

Interpretation and utilization of results. The author states that the test can be used in (1) diagnosis of language difficulties; (2) sectioning of classes; (3) measurement of

progress. However, its utility is lessened by reason of several considerations: (1) it has not been validated in accordance with scientific principles; (2) the norms given are inadequate; and (3) the reliability (.70 for college freshmen) is so low that the chances are only 4 to 1 that an obtained score is correct within 12 points. The difference between the lower and upper quartiles of college freshmen is 17 score points. However, the teacher can base tentative judgments upon the results obtained with the Cross English Test; and where time permits the giving of two — or better still three — forms, much of the unreliability indicated above will be eradicated.

Pressey Diagnostic Tests in English Composition

Description of the tests. The Pressey Diagnostic Tests in English Composition are four in number:

- (a) Capitalization (Pressey-Bowers)
- (b) Punctuation (Pressey-Ruhlen)
- (c) Grammar (Conkling-Pressey)
- (d) Sentence Structure (Conkling-Pressey)

Tests in this series are published in two forms, and are designed for use throughout junior and senior high school and college. All the tests are power tests, but the whole series can ordinarily be given in 42 minutes. Each test is published separately. Scoring is fairly rapid and is completely objective.

Interpretation and utilization of results. Grade norms (medians) based on many thousands of cases are given in the Manual of Directions, but they mean little in a test of this nature. The tests are designed to diagnose pupil difficulties, and the most fruitful procedure is the tabulation of errors in order to determine just where each pupil is weak and to indicate the greatest needs of the class. Drill work can then

be conducted accordingly. The Diagnostic Tests in English Composition are really an accompaniment to the *Student's Guide to Correctness in Written Work* prepared by S. L. Pressey and F. R. Conkling. This Guide is a 9-page booklet which contains "all the important rules for writing correct English" and is placed directly in the hands of the students. Preliminary statistical investigation showed that failure to observe the rules contained in this booklet accounted for 90 per cent of all the errors in capitalization, punctuation, grammar, and sentence structure. Each rule is very specific and is followed by concrete examples. A Teacher's Manual outlines desirable remedial procedures in connection with the use of the *Student's Guide* and the four diagnostic tests.

Validity and reliability. The items included in the Diagnostic Tests in English Composition comprise the principal errors made in these branches of English study. Data on the reliability are given in Table 17.

TABLE 17

DATA ON THE RELIABILITY OF THE DIAGNOSTIC TESTS IN ENGLISH COMPOSITION, FORM 1 (PRESSEY ET AL.). (1) CAPITALIZATION, (2) PUNCTUATION, (3) GRAMMAR, (4) SENTENCE STRUCTURE

r_{11}	r_{12}	N	S.D. events	S.D. odds	P.E. score	P.E. ∞ .1	$\frac{\text{P.E. score}}{\text{S.D.}}$	$\frac{\text{P.E. } \infty.1}{\text{S.D.}}$	NATURE OF GROUP	
(1)	.65	.79	99	2.0	2.3	.9	.6	.39	.28	Gr. IX, H. S.
(2)	.64	.78	99	3.2	2.8	1.2	.8	.40	.28	Gr. IX, H. S.
(3)	.83	.90	99	3.1	3.2	.9	.6	.27	.20	Gr. IX, H. S.
(4)	.58	.73	99	2.0	2.3	.9	.6	.43	.29	Gr. IX, H. S.

Wakefield Diagnostic English Test

The Wakefield Diagnostic English Test is really only diagnostic of the student's knowledge of formal terminology; it affords a measure of the phases of grammar least stressed in present-day methods.

Tressler English Minimum Essentials Test

Description of the test. The Tressler English Minimum Essentials Test is published in three forms, each comprising the following parts:

- (1) Grammatical Correctness (20 items)
- (2) Vocabulary (15 items)
- (3) Punctuation and Capitalization (6 items)
- (4) The Sentence and Its Parts (10 items)
- (5) Sentence Sense (10 items)
- (6) Inflection and Accent (10 items)
- (7) Spelling (15 items)

Each pupil is given as long as he needs for the test. In the high school grades slow pupils require from 40 to 50 minutes. Scoring is completely objective. Norms are given in the Manual of Directions, but the use of the test is intended to be diagnostic. Its utility in diagnosis is restricted because of the small sampling accorded each of the divisions of the test.

LANGUAGE-GRAMMAR TEST INTERCORRELATIONS

The authors have obtained intercorrelations among the various tests in order to show the extent to which different tests are measuring the same functions. These correlations are given in Table 18.

TABLE 18

	<i>r</i>	<i>N</i>
Briggs English Form Test ¹ vs. Kirby Grammar Test (Sentences)56	80
Briggs English Form Test vs. Kirby Grammar Test (Principles)63	80
Charters Diagnostic Grammar Test ¹ vs. Kirby Grammar Test (Sentences)48	80
Charters Diagnostic Grammar Test ¹ vs. Kirby Grammar Test (Principles)62	80
Pressey-Bowers Capitalization vs. Conkling-Pressey Grammar	.44	99
Pressey-Bowers Capitalization vs. Pressey-Ruhlen Punctuation47	99
Conkling-Pressey Sentence Structure vs. Pressey-Bowers Capitalization40	99
Conkling-Pressey Grammar vs. Pressey-Ruhlen Punctuation66	99
Conkling-Pressey Grammar vs. Conkling-Pressey Sentence Structure63	99
Conkling-Pressey Sentence Structure vs. Pressey-Ruhlen Punctuation55	99

Iowa Placement Examinations²

Two English examinations are included in the Iowa Placement series: English Aptitude, Revised (2 forms), and English Training, Revised (2 forms). The latter especially is primarily a language-grammar test. These examinations comprise the following material:

English Aptitude, EA-1, Revised

In Part 1 (20 items), the student is given a rule taken from a textbook used at the University of Iowa, together with samples of the applications of the rule. The test measures his ability to comprehend and apply the rule. In Part 2 (10 items) a passage of compact material is quoted from a

¹ See Chapter XIII.

² See Chapter XI.

college textbook, and the accuracy of the student's knowledge is measured. Part 3 (15 items) is an adaptation of the Iowa Comprehension Test method to material in English. Special emphasis is placed on the ideas gleaned from the passage rather than upon the factual content. Part 4 (20 items) measures composition ability through reference to an artificially devised composition printed on the page. The total working time of EA-1, Revised, is 43 minutes.

English Training, ET-1, Revised

In Part 1, 25 misspelled words are presented in a total list of 75 words, and the student is asked to write the misspelled words correctly. These words have been carefully drawn from the upper portion of the Iowa Spelling Scales devised by E. J. Ashbaugh. The relative difficulties of the different words as determined in that study have been recorded, and subsequent forms in this part of the test will be made of equal difficulty. Part 2 (60 items) is a test of punctuation, and is really more than this. It measures the student's knowledge of sentence structure. These items and those in the next two parts of this test were selected in accordance with the experience of college teachers of English. Part 3 (60 items) is a test in English grammar especially referring to colloquialisms, barbarisms, and common errors of oral English. Part 4 (45 items) measures the student's ability to distinguish between clear, emphatic sentences and those which are weak, confused, or ridiculous. The total working time of ET-1, Revised, is 40 minutes.

Remedial procedures in language and grammar. Teaching devices for improving understanding of rules of grammar and inculcating better language habits have already been touched upon. The first need is discovery of fundamental deficiencies in training. Where difficulty is encountered in improving the work of a student, his IQ should be determined

and he should be tested particularly for hearing and speech defects. Often the English work of the classroom will be heavily discounted by adverse home conditions and the influence of the pupil's companions; in such cases the pupil must be subjected to considerable over-learning.

II. SPELLING TESTS

Introduction. Why spelling remains a topic of some importance in high school work is concisely stated by the authors of the Seven S Spelling Scales:

The chief reason why tests have more frequently been standardized for the elementary grades than for the secondary school is that as education advances its purposes become more complex and the results consequently more difficult to measure. In spelling, however, the purpose is practically the same for the lower and for the higher grades, and there is need so long as pupils are in school to ascertain the relative proficiency of individuals, of classes, of schools, and of whole systems, by ages and by sex.¹

But measurement of spelling in high school involves factors which are not encountered in the lower grades. Ordinarily the pupil's vocabulary has been greatly extended, and heavily weighted in the direction of his special studies. Thorndike² has recently estimated the average vocabulary of high school students as ranging from 10,000 to 11,500 words. This estimate includes only words which the pupil can define, and omits the common tense, number, and degree inflections. The spelling lists briefly mentioned here are expected to supplement somewhat the lists which have grown out of the studies of Ayres, Horn, Ashbaugh, and others. Proper names and technical terms peculiar to certain sub-

¹ "Sixteen Spelling Scales." *Teachers College Bulletin*, Twelfth Series, No. 19 (May 21, 1921), page 1.

² Thorndike, E. L., "The Vocabularies of School Pupils." *Contributions to Education*, Vol. I, Chapter VII (World Book Company, 1924).

jects are not included in the Seven S and Monroe lists, but will be found in the fifteen lists compiled by Luella C. Pressey. Teachers of mathematics, science, geography, etc., should bear in mind that familiarity with the connotation of a technical term implies ability to pronounce and spell it correctly and to use it properly in sentences.

Sixteen Spelling Scales Standardized in Sentences for Secondary Schools (Seven S Spelling Scales)

Description of the scales. The Seven S Spelling Scales were developed at Teachers College, Columbia University, to meet the general needs of high school teachers. The scales consist of sixteen lists of sentences in which the significant words are embedded. The first twelve lists are of equal difficulty (on the basis of per cent of misspellings); likewise the next four are equal, but more difficult than the first twelve. In effect, the scales constitute two tests: the first of twelve equivalent forms, and the second of four equivalent forms. The words were drawn from the second and third thousand most commonly used words as determined by various studies. Ayres's list of the first thousand words and special forms of words were eliminated as beside the purpose of the scales. The social value of many of the words drawn from the early studies is open to question.¹

Administration and scoring. Directions for giving the test are carefully outlined. It can be given to a large group at one time. The teacher is directed to "Read the sentence aloud to the pupils; then pronounce the italicized word to denote which word they are to spell. Read the sentence and repeat the word again." To secure a spelling score of satis-

¹ See, for example, Morton, R. L., "The Reliability of Measurements in Spelling." *Journal of Educational Method*, Vol. III (April, 1924), pages 321-328.

factory reliability, it is recommended that at least forty words (two lists) be used for individual testing, and twenty words for testing a class.

Norms. Table 19 gives norms for February testing.

TABLE 19

GRADE	PER CENT CORRECT	
	Lists I-XII	Lists XIII-XVI
VII	65.90	34.76
VIII	73.77	45.03
IX	80.00	53.91
X	85.05	61.48
XI	88.67	67.08
XII	91.25	72.14

Standards for any other month are readily obtained by interpolation. Count ten months to the school year, and allow for each month one tenth of the difference between the standards between which the particular month occurs.

OTHER TESTS IN SPELLING

Monroe Timed Sentence Spelling Tests, III. This test includes a section for high school use. The sentences are dictated at a fixed rate which was determined as suitable for the pupils. No emphasis is placed on the significant words. To the pupils the test is simply a dictation exercise. Thus attention is not directed to the question of spelling, the assumption being that this is the more natural situation. For Grades IX to XII the test words were chosen from Columns S, T, and U of the Ayres Spelling Scale. As Ayres predicted, words embedded in sentences in this manner resulted in standard scores somewhat lower than those he furnished. Since Monroe's norms for this test range from 86 to 96 (per

cent) for the high school grades, the test is really too easy to measure spelling ability in these grades.

Remedial procedures in spelling. It is doubtless true that for the average high school English teacher the problem of spelling must remain in the background. In many cases he is confronted with pupils whose defects can be traced to indifferent instruction in the lower grades. Their errors are brought to light chiefly in student themes. Hudelson's questionnaire¹ showed that most teachers indicate mechanical errors in themes, but the majority do not correct them. Since correcting is left to the pupil, he should be taught to meet this responsibility. Good spelling habits are essential. There is certainly little transfer of ability from one word to another; each misspelled word calls for a specific learning task. Tests will show whether the fundamental spellings have been acquired, and diagnostic charting will furnish a guide as to the chief sources of errors. In the higher grades especially, the pupil can no longer expect to be drilled in spelling. The more difficult words will be assimilated only through the building up of careful reading and dictionary habits. The following aids in extending a meaning vocabulary are proposed by the National Committee on Reading²:

- (1) Rapid growth of vocabulary through actual experience and wide reading. Special attention to words and idioms significant in new subjects.
- (2) Attention to words and groups of words in context whose value will be increased because of intensive work done with them in composition, grammar, and foreign language study. This should be a carrying over of training and should illuminate the context.

¹ Hudelson, Earl, *The Twenty-second Yearbook of the National Society for the Study of Education*, Vol. XXII, Part I, pages 1-3 (Public School Publishing Company, 1923).

² "Report of the National Committee on Reading." *The Twenty-fourth Yearbook of the National Society for the Study of Education*, Vol. XXIV, Part I, pages 95-96 (Public School Publishing Company, 1925).

- (3) Intensive study of carefully selected words to show wealth of English language and the values of words in expressing shades of meaning.
- (4) Training in making and interpreting definitions, usually depending upon synonyms and illustrative sentences.
- (5) Training in judging relative values of words in context, so that dictionaries and other helps may be sensibly used without over-emphasis on detail.
- (6) Exercises in classification of words as to thought; arranging word lists under appropriate headings; making lists of synonyms and synonymous expressions, of antonyms, of words with common roots, prefixes, and suffixes.
- (7) Training in knowing and using all the resources of the dictionary.
- (8) Testing vocabulary growth by both informal and standard tests.

III. READING TESTS

Introduction. The place of reading in high school has been investigated by the National Committee on Reading, of the National Society for the Study of Education. The judgments of this Committee, composed of leading authorities on Reading, are contained in the *Twenty-fourth Yearbook*.¹

The high school work should cover the "period of refinement of specific reading attitudes and habits, and tastes. During this period, reading and study habits are refined in each content subject as well as in the literature period. Wholesome interests in reading, the habit of reading current events and books and magazines of real worth, the sources of different types of reading materials, and standards of selection are emphasized" (page 25). It is pointed out that other phases of reading which may be considered the chief purposes in the lower grades are continued throughout the high school in the interests of increasing perfection.

¹ "Report of the National Committee on Reading." *The Twenty-fourth Yearbook of the National Society for the Study of Education*, Vol. XXIV, Part I (Public School Publishing Company, 1925).

The same report lists two distinguishing characteristics of the junior and senior high school period: (1) refinement and perfection of attitudes, habits, and tastes previously developed; (2) emphasis on conscious learning; the pupils now deliberately study and seek to improve their own reading habits. These characteristics are to be taken into account by directing the work toward six specific aims, as follows: (1) extension of the experiences and "intellectual apprehension" of pupils; (2) further development of interests and tastes which will "direct and inspire the present and future life of the reader and provide for the wholesome use of leisure time"; (3) stimulation of habits of intelligent interpretation; (4) provision for individual and group instruction in fundamental reading habits; (5) further development of oral reading, particularly of literary and dramatic selections; (6) development of skill in the use of books and library privileges.

When the teacher turns to the tests developed for the measurement of reading in high school, he discovers that they obviously have not been validated in accordance with such aims. They measure fairly well general comprehension of reading material, but they are diagnostic only to a very limited degree, and the remedial measures which should be undertaken as a result of the testing are not often clear. The teacher can only meet the aims of the better type of English teaching in the high school by resourcefulness in developing his own measures of pupil diagnosis and pupil progress.

Haggerty Reading Examination, Sigma 3

Description of the examination. The Haggerty Reading Examination, Sigma 3, is designed for Grades VI to XII. Two forms are available, each consisting of three tests: Vocabulary, Sentence Reading, and Paragraph Reading, all

three comprising an 8-page folder. Each test is preceded by a fore-exercise.

What the examination measures. The examination measures reading ability or reading comprehension in general; it is more than an intelligence test, for no attempt is made to differentiate between inherited capacity and training. It is probable that Test 2 (Sentence Reading) is essentially a vocabulary test as well as Test 1. Test 3 measures the ability to secure exact meanings from rather compact paragraphs.

Administration and scoring. Exact directions for giving the test are found in the Manual of Directions. Two methods are outlined. Method A allows the pupils ample time to do the fore-exercises and then times the pupils on only the test proper. In Method B the fore-exercises are considered part of the test and the pupils are timed from the moment they begin to read the directions. This method involves comprehension of material different from that in the test, and introduces an undesirable, variable element in the procedure. Hence it is recommended that teachers follow Method A. The time allowance for the three tests in the examination is 5, 3, and 20 minutes, respectively. The whole test can be given readily in one class-hour. Scoring is completely objective and fairly rapid.

Interpretation and utilization of results. The results of the Haggerty Reading tests may be used for sectioning pupils on the basis of reading ability, and for measuring pupil progress in reading. The norms given do not represent a high degree of achievement and should be thought of rather as minimum standards for average students. They can also be used for diagnosis of pupil difficulties in vocabulary, sentences, and paragraphs. The tests show a fair amount of discrimination between grades and between chronological age increments of one year.

Validity and reliability. The reading materials which form the basis of these examinations were an outgrowth of surveys of the public schools in St. Paul, Minnesota, and in North Carolina. Items were originally drawn from school readers and textbooks in United States History. The author gives the reliability of Sigma 3 as .89 for a range of talent including Grades V-C to VIII-A. (See also Table 57.)

Norms. Age and grade norms (Grades V to XII and ages 10 to 20) are given in the Manual of Directions. In addition, a considerable amount of supplementary data in connection with norms and reliability were accumulated in the Rural School Survey of New York State conducted by Haggerty.¹

Thorndike-McCall Reading Scale

Description of the scale. The Thorndike-McCall Reading Scale is issued in ten equivalent forms. The scale can be used from Grade III through high school. It is similar to Thorndike's Scale Alpha 2 for Measuring the Understanding of Sentences, but the latter is not suitable for high school use. McCall (in *How to Measure in Education*, page 272) describes the method of devising the scale:

Selections of prose and poetry were made which were brief, which gradually varied in difficulty from very easy to very difficult, which were fairly representative of reading material in school and out, which were reasonably free from technical terms, and which were equally fair to rural and urban children.

Questions were formulated which could be answered from or inferred from or were related to the reading selections, which would yield brief, scorable answers, which were unambiguous, whose difficulty approximated that of the selection, which were independent either in wording or difficulty of any preceding or succeeding

¹ Haggerty, M. E., *Rural School Survey of New York State Educational Achievement* (Joint Committee on Rural Schools, Ithaca, New York, 1922; 223 pages).

questions, and which were numerous enough to make up one test of the desired length.

The questions were then arranged in order of their difficulty for school children. A fore-exercise incorporates the directions which the pupils are to follow.

What the scale measures. It is difficult to tell exactly what a reading test measures, for reading ability must be thought of as an organized group of skills, various parts of which are often erroneously singled out as a complete measure of reading. It is clear, however, that this scale is a measure of power, of reading comprehension, and not of reading rate. The time is limited, but is ample for any pupil who can grasp the material at all. Gates¹ points out that it is "probably the only test measuring a certain type of power in comprehension, unaffected by the mechanical factors of reading." He indicates, too, that it should not be used as a measure of the effectiveness of instruction, since it is "little subject to improvement through specific practice." More experimentation is needed on this point, especially in connection with its relation to the broader concept of general intelligence. That pupils do reach higher scores in the test each year is known, but how much of this is due to increased maturity and how much to more instruction or better instruction has not been determined.

Administration and scoring. The Thorndike-McCall Scale is readily administered. The pupils learn what is wanted in going through the fore-exercise. The actual working time is 30 minutes. The scoring key consists of answers accepted as correct, together with some of the more common types of incorrect response. Scoring is tedious, but is facilitated by practice. The teacher should adhere as strictly as possible

¹ Gates, A. I., "Experimental and Statistical Study of Reading and Reading Tests." *Journal of Educational Psychology*, Vol. XII (September, 1921), pages 303-314.

to the seven specific rules given in the Manual of Directions. The first score obtained is simply the number right. Tables should then be prepared from data given in the Manual of Directions, leading to the following measures: *T*-Scores, Reading Age, and Reading Quotient. These derived scores enable comparisons of pupil-standing from test to test in different fields. Conversion tables are given in the Manual of Directions for the Thorndike-McCall Scale, and the norms are based upon them. Another table converts the *T*-Scores into a Reading Age. The latter is expressed in months and may be divided by the pupil's chronological age to obtain his Reading Quotient. As with the IQ, the "normal" child will have a quotient of 100, meaning here that he possesses normal reading ability. Inferior and superior children with respect to reading will be found proportionately below and above 100.

Interpretation and utilization of results. So much of statistical nature has been written about this scale, that there is a real danger that the teacher, not interested in such discussions *per se*, may find the issues obscured. In brief, the Thorndike-McCall Reading Scale is not as reliable as many of the later reading tests — e.g., the Haggerty Reading Examination and the Stanford Reading Examination — and do not have the diagnostic power of these later tests. (See Tables 20 and 57.) Reading a paragraph, then finding the right answer to questions based on the reading (the paragraph remaining in view), is to some extent a matching exercise in silent reading; it disregards the mechanics of the process, takes no account of speed, and gives only a partial measure of the material the pupil really grasped. These factors may be regarded, of course, as outside the scope of the scales, but they are not outside the scope of the teaching of reading.

These scales serve certain useful purposes, however. They enable the teacher to locate the "capable but slow" pupil

in reading, and to place him approximately in the proper grade. Other reading tests usually fail to do this, as speed is ordinarily a consideration. The general standing of a class or a school system can be ascertained, for extensive norms are available. The large number of equivalent forms permits giving the test as frequently as desired.

Validity and reliability. The method of validation of the Thorndike-McCall Reading Scale has already been indicated. Data on reliability are given by Gates and McCall. (See References.) Additional data are given in Table 20. (See also Table 57, Chapter XIII.)

TABLE 20

DATA ON THE RELIABILITY OF THE THORNDIKE-MCCALL READING SCALE

r	N	S.D.	P.E.-score	P.E. ∞ .1	P.E.-score S.D.	P.E. ∞ .1 S.D.	NATURE OF GROUP
.58	27	7.0 (T 's)	3.1 (T 's)	2.3 (T 's)	.44	.33	Grade V
.59	32	3.2	1.4	1.1	.44	.33	Grade V
.75	154	4.5	1.5	1.3	.34	.29	Grades IV-VIII

Norms. Age and grade norms on over 10,000 cases are given in the Manual of Directions. Tables are given also for translating scores into T -Scores and Reading Age.

Van Wagenen Reading Scales — English Literature

Description of the scales. The Van Wagenen Reading Scales in English Literature are part of a series which includes reading tests in history, general science, and English literature. Three scales, A, B, and C, are available in English literature. Each consists of series of paragraphs, 15 in number, followed by statements regarding the content. Paragraph A (a fore-exercise) is reproduced below:

DIRECTIONS

Read Paragraph A carefully. Then read the statements below it and put a check mark (✓) on the dotted line in front of each statement which contains an idea that is in the paragraph or that can be derived from it. The first statement is already checked as it should be. The paragraph and statements may be re-read as often as it is necessary. (Paragraph A is a sample for practice.)

Paragraph A. Our breakfast consisted of what the squire denominated true old English fare. He indulged in some bitter lamentations over modern breakfasts of tea and toast, which he censured as among the causes of modern effeminacy and weak nerves, and the decline of old English heartiness; and though he admitted them to his table to suit the palates of his guests, yet there was a brave display of cold meats, wine, and ale on the sideboard.

-✓.....1. The squire did not approve of breakfasts of tea and toast.
-2. The squire was considerate of the habits of his guests.
-3. The breakfasts of the peasants consisted of meats and wine or ale.
-4. The squire believed that English life had lost some of its vigor and wholesomeness.
-5. The squire's breakfast consisted only of tea and toast.

What the scales measure. In going through the scales the pupil must resist falling into error, and must depend upon inference to a considerable extent. The test is much more than a matching test, as intelligence enters into it markedly. This type of scale abstracts definitely the comprehension of the material from its general "appreciation." Scoring is completely objective, although somewhat complicated by a weighting system. The end result is a numerical score.

Interpretation and utilization of results. Class medians, 25-percentiles, and 75-percentiles can be compared to the tentative standards of achievement given in the Class Record Sheet. An individual record card permits ready transference of the scores into mental ages (derived from comparison of about 500 cases given in the English Literature Scale and the Terman Group Test of Mental Ability). The Van Wagenen Reading Scales can be used to advantage where comprehen-

sion of a particular type of reading should be measured. Ordinarily the results should be considered as supplementary to those of a standard intelligence test.

Monroe Standardized Silent Reading Tests, III

Description of the test. Test III of the Monroe Standardized Silent Reading Tests is designed for Grades IX, X, XI, and XII. Two forms are available. After each paragraph a single significant question is asked, the type response varying from 2-response to recall, but being fairly objective in all cases. A fore-exercise familiarizes the pupils with the procedure. The test requires only 5 minutes' working time.

What the test measures. The test measures the ability to read paragraphs common to the range of experience of high school pupils, and to draw the essential meaning from these paragraphs. It also gives a measure of the rate of reading under these conditions. The chief deficiency of the test is its brevity. It is too short to be utilized for individual measurement, and the standards fail to mark class progress in the high school range of talent. Reading is so much at the heart of all pupil progress that a 5-minute test cannot serve to cover even one phase of it adequately. Table 57, Chapter XIII, gives data on the reliability of the Monroe test.

Pressey Technical Vocabularies of the Public School Subjects

These vocabularies consist of fifteen lists, published separately; viz.:

(1) Grammar and Composition (English, French, Latin, German)	(5) History
	(6) General Science
(2) Literature	(7) Biology
(3) Arithmetic	(8) Chemistry
(4) Algebra and Geometry	(9) Geography
	(10) Physics

(11) Physiology	(14) Art
(12) Home Economics	(15) Music
(13) Manual Training	

These lists were prepared by Luella C. Pressey in accordance with the ratings of teachers on the importance of technical words appearing in common textbooks. The author of the vocabulary lists recommends that each pupil go through the list for the subject being taught and systematically master every word. She states also that

these lists are more than mere lists of words; they are catalogues of the important concepts in each subject. So the lists may very profitably be used as "diagnostic tests" in determining the specific weaknesses, of a class or an individual pupil, in a subject. That is, if the teacher will find out the terms which a pupil does not know, in the vocabulary for the subject in question, she will have a very valuable indication as to the topics on which that pupil needs help. The lists can be of great service in thus diagnosing weaknesses, and guiding the teacher in her remedial instruction.

Holley Sentence Vocabulary Scale

Series 3 B for Grades VII to XII consists of the last 70 words in the vocabulary test of the Stanford Revision of the Binet-Simon Scale, presented to the pupil in a 4-response situation. This test can also be used as a rough measure of intelligence. The reliability of the Holley Sentence Vocabulary Scale was found to be .82 on 140 high school students, all grades, and the probable error of a score 2.4. Terman (in *The Measurement of Intelligence*) gives a complete account of the validation and utility of the words in the scale.

Remedial procedures in reading. When the teacher has found out all he can about the pupils' reading ability, his task has only begun. Specific suggestions have already been made, but there remains the broader problem of linking technique with the aims of the English course. The 1917

Report of the National Joint Committee on English¹ sets forth among the aims of an English course the following kinds of reading ability:

- (1) Cursory reading, to cover a great deal of ground, getting quickly at essentials.
- (2) Careful reading, to master the book, with exact understanding of its meaning and implications.
- (3) Consultation, to trace quickly and accurately a particular fact by means of indexes, guides, and reference books.

In these the pupil should be skillful, and should know when to use each. Teachers of English will no doubt agree that the habits implied in types (1) and (2) should be thoroughly ingrained before the high school age; while the third type demands library facilities not always available even in high schools. The first aim listed above is often lost sight of by the teacher himself. "Skipping" is a practice confused with rapid, intensive "skimming." As a matter of fact, a good word might well be said for both. In real life bulky, wordy newspapers and magazines thrust themselves upon the attention and the time of the schoolboy; it is a worth-while accomplishment to develop a "scent" for the essentials of the discussion. Standard tests have concentrated on comprehension of reading material, and their diagnostic utility is not great. It is recommended that English teachers develop informal, objective examinations covering the other phases. For a complete tabulation of reading deficiencies, together with corresponding diagnosis and remedial procedures, the teacher of English is referred to Part I of *The Twenty-fourth*

¹ Hoscic, J. F. (Chairman), *Reorganization of English in Secondary Schools*. Report by the National Joint Committee on English, Representing the Commission on the Reorganization of Secondary Education of the National Education Association and the National Council of English. Bulletin No. 2 (1917), page 32 (Department of the Interior, Bureau of Education, Washington, D. C.).

Yearbook of the National Society for the Study of Education, especially Chapters X and XI.

IV. ENGLISH COMPOSITION SCALES

Introduction. English composition is one of the most difficult of all subjects to measure objectively, and the principles of its effective measurement are perhaps the least understood in educational testing. The authors of the various scales generally recognize acutely the defects and the limitations of their attempts, but subsequent studies based on the results of measurement with the scales are likely to become extravagant in their conclusions. The purpose of teaching English composition is "to enable the pupil to speak and write correctly, convincingly, and interestingly. The first step toward efficiency in the use of language is the cultivation of earnestness and sincerity; the second is the development of accuracy and correctness; the third is the arousing of individuality and artistic consciousness." (1917 Report of the National Joint Committee on English.)

If we accept this as a fair statement, it becomes evident why the measurement of these accomplishments through the medium of a standardized scale is a most herculean task. Hudelson,¹ in the course of experimentations on this problem, obtained from high schools 165 responses which throw some light on what English teachers actually do throughout the country. His chief conclusions may be summarized as follows:

- (1) Teachers are trying to secure *general merit* in the compositions.
- (2) Rhetorical principles receive most emphasis in composition teaching.

¹ Hudelson, Earl, *The Twenty-second Yearbook of the National Society for the Study of Education*, Vol. XXII, Part I, page 15 (Public School Publishing Company, 1923).

- (3) Practically all compositions are filed either in the classroom or in individual notebooks. These files are apparently not consulted afterward.
- (4) Errors and weaknesses are seldom corrected by the teacher, except in matters of taste, questions demanding nice judgment, or cases involving unfamiliar principles.
- (5) The aims of composition teaching as set forth in United States *Bureau of Education Bulletin*, 1917, No. 2, are widely known, generally accepted, and deliberately departed from only in a few and comparatively minor details.

What the composition scales measure. It is fairly definite knowledge what the scales should attempt to measure; the real difficulty lies in hitting upon a good method of testing. Scales have taken over for their particular field the measurement of *general merit* in composition. They do not attempt, for the most part, to analyze this general merit. The unit of measurement is usually that amount of difference in general merit which is noticed by 75 per cent of a group of trained judges. That is, if 50 per cent of the judges believe composition A better than B, and 50 per cent that B is better than A, then A is said to possess the same merit as B. An increase in the per cent voting for composition A increases the likelihood of its being really superior, and it is considered just one unit superior to B when it is deemed superior by 75 per cent of the judges.

The scales themselves consist of a series of graded compositions, chiefly bona fide pupil themes, advancing from zero or very low merit to very high merit. For the teacher the use of the scales becomes a matter of careful matching. A theme is "moved along" a scale until it appears to be of the same or nearly the same merit as one in the printed series.

The theme is then given the merit-grade of the standard composition, which is printed on the scale in each case. Certain modifications of this general procedure will be indicated in the paragraph assigned to each of the more widely used scales. Variations of importance are found in (1) the method of securing the sample of the pupil's composition work; (2) length of model compositions and fineness of steps between them; (3) nature of the final score; (4) reliability and validity of interpretations of composition performance with respect to the pupil and to the school system.

The Hillegas Scale of 1912 is the starting point for composition scales now in use. It was devised to measure general merit and it employed the per-cent-of-judges-agreeing for defining its units. The Hillegas Scale has been revised and extended, so that it now has historical interest only. The Thorndike Extension of the Hillegas Scale for the Measurement of Quality in English Composition by Young People includes 29 compositions and 15 units, ranging in value from 0 to 95. Near the middle of the scale several equivalent compositions are given for each step. The Nassau County Supplement to the Hillegas Scale (by M. R. Trabue) embodies several notable improvements, but it in turn is less useful than the Hudelson Scale.

The Hudelson English Composition Scale is also an outgrowth of the Hillegas Scale. Noteworthy features are the uniformity and fineness of intervals and the equivalence of its values to those obtained by the Thorndike Extension and the Nassau County Supplement. For the average teacher the fineness of steps is illusory, since only highly trained judges can make use of the fractional intervals. Moreover, the scale does not run down to zero, the higher samples are artificial, and there is only one example for each merit-value. The reliability of a single teacher's rating of pupils' papers by the Hudelson English Composition Scale is not always

clear in published reports, since figures are often given based on pooled estimates of a number of judges. The English teacher must bear in mind that to attain the accuracy of a pooled estimate his judgment must similarly be corroborated by that of others. Hudelson recommends that three judges rate the same themes, and this is especially desirable when important decisions or comparisons are to be made.

The Hudelson Typical Composition Ability Scale and the Hudelson Maximal Composition Ability Scale are new departures in composition measurement. Hudelson considers them as means toward certain definite ends; e.g., testing the results of teaching composition by various methods, individual pupil comparison and classification, and the supplying of an incentive to competition. The Maximal Composition Scale is intended especially for pupil classification, and the Typical Composition Scale for the grading of pupils, but both contribute to other aims. These scales are not teaching devices and should not be used as such. At the beginning or at the end of the semester or school year is the most effective time to employ them.

Administration of the Hudelson Composition Scales. Administration is reduced to a standard procedure for both scales. For measurement by means of Typical Composition the teacher reads a composition (provided with the scale), and the pupils write fifteen minutes, telling the same story in their own words. For measuring maximal composition ability the teacher simply announces one of the selected topics and the pupils write fifteen minutes on this assigned topic. In each case the resulting compositions are compared to their respective scales and a score is assigned to them by the usual sample-matching method. The scale ranges from 0 to 9 in gradations of one, and the teacher is expected to interpolate fractional values. For re-testing, various other topics have been selected which will also result in a measurement of

maximal composition ability. National standards for January testing are given in the Teacher's Handbook, a pamphlet provided with the scales. These norms may be used with Nassau, Hudelson, and Lewis scales.

OTHER COMPOSITION SCALES

The Lewis English Composition Scales are devised to measure the following kinds of letters: simple order letters, applications, narrative social letters, expository social letters, and narratives on a topic arbitrarily assigned. A monograph¹ gives complete information concerning the formation and use of these scales. Teachers of English interested in this phase of measurement should obtain a copy of the monograph and undertake experiments with the Lewis Scales. They were carefully constructed and are as reliable as the more general composition scales.

The Van Wagenen English Composition Scales seek to measure "general merit" in a more analytical way. Separate scales are provided for exposition, narration, and description. Each scale is graded with respect to Thought Content, Structure, and Mechanics, the scale values being kept separately. Compositions are scored by matching them with the scale samples with respect to the three elements of English given above. However, the three scores obtained are combined into a total score by a method "which gives weights to the three ratings in proportion to their importance." The unit of merit is the same as for the earlier scales, but the steps between samples are not uniform. As Van Wagenen demonstrates, this need not be a disadvantage. Rather detailed suggestions are made as to the points to be kept in mind in rating each element, and of course the in-

¹ Lewis, E. E., *Scales for Measuring Special Types of English Composition* (World Book Company, 1921; 142 pages).

fluence of factors grouped under an element not being rated at the time must be subjectively disregarded. Thus the whole procedure is somewhat confusing, but Van Wagenen states that a small amount of practice brings facility in using the scales. Sample themes are given (with detached true ratings) for practice. A certain amount of steadiness, and the elimination of constant errors, are required before one applies the scale to student themes.

Reliability of composition scales. Data obtained by the authors on the reliability of composition measurements are given in Table 21.

TABLE 21
DATA ON THE RELIABILITY OF COMPOSITION MEASUREMENTS

SCALE OR PROCEDURE	<i>r</i>	<i>N</i>	NATURE OF GROUP
(1) Teacher 1 <i>vs.</i> Teacher 2 (without the use of scale) .	.80	50	Grade IX, H.S. same pupils
(2) Van Wagenen — General Merit67	50	Grade IX, H.S. same pupils
(3) Van Wagenen — Structure .	.70	50	Grade IX, H.S. same pupils
(4) Van Wagenen — Thought .	.55	50	Grade IX, H.S. same pupils
(5) Van Wagenen — Mechanics .	.50	50	Grade IX, H.S. same pupils
(6) Willing Scale — Story Value ¹	.40	50	Grade IX, H.S. same pupils
(7) Willing Scale — Form Value ¹	.92	50	Grade IX, H.S. same pupils
(8) Grades (without scale) <i>vs.</i> Van Wagenen General Merit	.64	50	Grade IX, H.S. same pupils
(9) Grades (without scale) <i>vs.</i> Van Wagenen General Merit	.73	50	Grade IX, H.S. same pupils (Teacher 2)

Remedial procedures in English composition. The best way for a teacher to judge composition scales is to become familiar with their use. Specific remedial measures will then occur to the teacher. A few suggestions may be helpful:

¹ See Chapter XIII.

- (1) Scores which stand for "general merit" must not be invested with diagnostic powers, and such scores do not necessarily furnish a fair basis for comparisons as between individual pupils, grades, or school systems.
- (2) The scales are not teaching devices; they should only be employed at a general "checking-up" time.
- (3) A group of teachers should feel free to construct their own scales. Such scales might supplement the nationally known scales, as do objective examinations standard tests.
- (4) Whether or not a scale is employed, the teacher should adopt some systematic method of scoring and ranking pupil themes; and having found a distribution of innate capacity and of performance in composition, seek to correct not only mechanical defects but the more subtle difficulties of composition which confront the individual pupil.

Gainsburg,¹ in a rather sweeping criticism of present composition measurement methods, sets up the following "vital factors" in composition:

- (a) Impression (rather than expression)
- (b) Content (rather than phraseology and style)
- (c) Originality (rather than reproduction)

These factors, all will agree, are difficult to control, but they must be kept in mind if the English teacher is to escape formalizing the spirit of composition measurement. A composition, after all, is made to be *read*, not merely written. Is it clear, interesting, convincing, compelling, emotion-arousing, challenging to the reader? English teachers must answer this question somehow. Any scale or procedure

¹ Gainsburg, J. C., "Fundamental Issues in Evaluating Composition." *The Pedagogical Seminary* (March, 1924), pages 55-77.

which promises to contribute to a satisfactory solution should be given a thorough trial.

V. MISCELLANEOUS TESTS IN ENGLISH¹

Various tests have been devised to measure particular phases of English. Greene's Organization Test consists of a series of disarranged sentences which pupils are to arrange in logical sequence. The score combines power and rate. The specific ability measured is one sometimes incorporated in tests of intelligence. Abbott and Trabue's Exercises in Judging Poetry are published in two forms. The pupil is confronted with 13 sets of stanzas. Each set comprises the original poetic rendition and 3 stanzas derived from the original. These artificial versions represent mutilations of emotional quality, imagery, and meter, respectively. Whipple's High School and College Reading Test has for its object the determination of "how rapidly students in the high school and college are able to read and comprehend such material as they encounter in their daily work." It is adapted to mature readers and requires sustained attention. Tentative norms are given in the Manual of Directions. Logasa and McCoy, of the University High School, University of Chicago, have recently published, after two years' experimentation, preliminary forms of Seven Tests for Appreciation of Literature. The tests cover: Discovery of Theme, Reader Participation, Sensory Images, Comparisons, Rhythm, Trite and Fresh Expression, Standard of Taste. The working time for each test is from 12 to 19 minutes.

¹ See Chapter XI for the Columbia Research Bureau English Test.

Test Materials

LANGUAGE—GRAMMAR

Kirby Language and Grammar Test. By THOMAS J. KIRBY. Forms I and II, each form with directions, \$1.75 per 100. Bureau of Educational Research and Service, Extension Division, University of Iowa, Iowa City, Iowa.

Wilson Language Error Test. By G. M. WILSON. Examination Booklet, with Manual of Directions and Key, Percentile Graph, and Class Record, 80 cents per package of 25. Specimen set, 10 cents. World Book Company, Yonkers-on-Hudson, New York.

Cross English Test. By E. A. CROSS. Forms A, B, and C, each form, with Manual of Directions, Key, and Class Record, \$1.20 per package of 25. Specimen set, 25 cents. World Book Company, Yonkers-on-Hudson, New York.

Pressey Diagnostic Tests in English Composition. Form I (a) Capitalization, by S. L. PRESSEY and E. V. BOWERS, 75 cents per 100; (b) Punctuation, by S. L. PRESSEY and HELEN RUHLEN, 75 cents per 100; (c) Grammar, by F. R. CONKLING and S. L. PRESSEY, \$1.50 per 100; (d) Sentence Structure, by F. R. CONKLING and S. L. PRESSEY, \$1.50 per 100. Form II, equivalent tests by S. L. PRESSEY. Specimen set, including all four tests, 15 cents. Student's Guide to Correctness in Written Work, by S. L. PRESSEY and F. R. CONKLING, \$5.00 per 100. Teacher's Manual (for use with guide), 5 cents per copy. Public School Publishing Company, Bloomington, Illinois.

Tressler English Minimum Essentials Test. By J. C. TRESSLER. Forms A, B, and C, each 75 cents per package of 25. Specimen set, 10 cents. Public School Publishing Company, Bloomington, Illinois.

Iowa Placement Examinations. English Aptitude, EA-1, Revised, Forms A and B and English Training, ET-1, Revised, Forms A and B. Each form, \$3.50 per 100, with Manual of Directions and Scoring Key. Bureau of Educational Research and Service, University of Iowa, Iowa City, Iowa.

SPELLING AND VOCABULARY

Sixteen Spelling Scales Standardized in Sentences for Secondary Schools. (Seven S Spelling Scales.) Prepared under the direction of THOMAS H. BRIGGS and TRUMAN L. KELLEY. For examiner: One bulletin of title given above, 40 cents. (Scales not published separately.) Bureau of Publications, Teachers College, Columbia University, New York City.

Monroe Timed Sentence Spelling Tests, III. By W. S. MONROE, \$4.00 per 100. Public School Publishing Company, Bloomington, Illinois.

Pressey Technical Vocabularies of the Public School Subjects. By L. C. PRESSEY. Prices range from 40 cents to \$1.50 per package of 35. Specimen for any list, 5 cents; complete specimen (15 lists), 75 cents. Public School Publishing Company, Bloomington, Illinois.

Holley's Sentence Vocabulary Scale. By C. E. HOLLEY. 80 cents per 100. Specimen set, 6 cents. Public School Publishing Company, Bloomington, Illinois.

READING

Haggerty Reading Examination, Sigma 3. By M. E. HAGGERTY. Forms A and B, each form with Key and Class Record Sheet, \$1.10 per package of 25. Manual of Directions, 25 cents. Specimen set, 45 cents. World Book Company, Yonkers-on-Hudson, New York.

Thorndike-McCall Reading Scale for the Understanding of Sentences. By E. L. THORNDIKE and W. A. McCALL. Ten forms. For examiner: Manual of Directions and Class Record Sheet supplied with each order. For pupil: Test form, \$2.00 per 100. Specimen set (one copy of Form 1 and directions bulletin), 10 cents. Bureau of Publications, Teachers College, Columbia University, New York.

Van Wagenen's Reading Scales — English Literature. By M. J. VAN WAGENEN. Scales A, B, and C, each scale \$3.00 per 100. Specimen set of all Reading Scales, 20 cents. Public School Publishing Company, Bloomington, Illinois.

Monroe Standardized Silent Reading Tests, III. By W. S. MONROE. Forms 1 and 2, each form \$1.00 per 100. Specimen set, 6 cents. Public School Publishing Company, Bloomington, Illinois.

Whipple's High School and College Reading Test. By G. M. WHIPPLE. Forms A and B, each \$3.00 per 100. Specimen set, 15 cents. Public School Publishing Company, Bloomington, Illinois.

COMPOSITION SCALES

Hudelson English Composition Scale. By EARL HUDELSON. 25 cents net per copy. Directions are included. World Book Company, Yonkers-on-Hudson, New York.

Hudelson's Typical Composition Ability Scale. By EARL HUDELSON. \$1.00 per package of 25; single copy, 10 cents. Teacher's Handbook, 10 cents per copy. Specimen set, 20 cents. Public School Publishing Company, Bloomington, Illinois.

Lewis Scales for Measuring Special Types of English Composition. By E. E. LEWIS. A School Efficiency Monograph, paper bound, \$1.36. Five Scales only, 8 pages. 25 cents net. World Book Company, Yonkers-on-Hudson, New York.

Van Wagenen English Composition Scales. By M. J. VAN WAGENEN. 25 cents net per copy. World Book Company, Yonkers-on-Hudson, New York.

MISCELLANEOUS

Abbott-Trabue's Scales for the Appreciation of Poetry. By A. ABBOTT and M. R. TRABUE. Series X and Y, each 5 cents per copy. Manual of Directions, 25 cents. Bureau of Publications, Teachers College, Columbia University, New York.

Logasa-McCoy's Seven Tests for Appreciation of Literature. By HANNAH LOGASA and MARTHA J. MCCOY. For particulars write Hannah Logasa, University High School, University of Chicago, Chicago, Illinois. 10 cents for one set of seven tests. Public School Publishing Company, Bloomington, Illinois.

References

ABBOTT, A., and TRABUE, M. R. "A Measure of Ability to Judge Poetry." *Teachers College Bulletin*, Fourteenth Series, No. 2 (September 23, 1922); 19 pages.

ASHBAUGH, E. J. "Senior High School English as Revealed by a Standard Test (The Tressler Test)." *Journal of Educational Research*, Vol. XIII, No. 4 (April, 1926), pages 249-258.

DICKINSON, C. E. "A Study of the Relation of Reading Ability and Scholastic Achievement." *School Review*, Vol. XXXIII, No. 8 (1925), pages 616-626.

DOLCH, E. W. "The Measurement of High School English." *Journal of Educational Research*, Vol. IV (November, 1921), pages 279-286.

— "More Accurate Use of Composition Scales." *The English Journal*, Vol. XI (November, 1922), pages 536-544.

GAINSBURG, J. C. "Fundamental Issues in Evaluating Composition." *The Pedagogical Seminary*, Vol. XXXI (March, 1924), pages 55-77.

GATES, A. I. "Experimental and Statistical Study of Reading and Reading Tests." *Journal of Educational Psychology*, Vol. XII (September, 1921), pages 303-314; (October, 1921), pages 378-391; (November, 1921), pages 445-464.

GOOD, C. V. "An Analysis of the Reading Recommendations Included in School Surveys." *Educational Administration and Supervision* (December, 1925), pages 577-587.

GRAY, C. T. *Deficiencies in Reading Ability*. D. C. Heath & Co., New York; 1922. 420 pages.

GRAY, W. S. "The Value of Informal Tests of Reading Accomplishment." *Journal of Educational Research*, Vol. I (January, 1920), pages 103-111.

HAGGERTY, M. E. *Rural School Survey of New York State Educational Achievement*. Joint Committee on Rural Schools, Ithaca, New York; 1922. 223 pages.

HILLEGAS, M. B. "A Scale for the Measurement of Quality in English Composition for Young People." *Teachers College Record*, Vol. XII (September, 1912).

HOSIC, J. F. (Chairman). *Reorganization of English in Secondary Schools*. Report by the National Joint Committee on English, Representing the Commission on the Reorganization of Secondary Education of the National Education Association and the National Council of English. Bulletin No. 2 (1917); 181 pages. Department of the Interior, Bureau of Education, Washington, D. C.

— "Some Needed Investigations in the Field of English." *Contributions to Education*, Vol. I, Chapter V. World Book Company, Yonkers-on-Hudson, New York; 1924.

HEDDELSON, EARL. "English Composition: Its Aims, Methods, and Measurement." *The Twenty-second Yearbook of the National Society for the Study of Education*, Part I. Public School Publishing Company, Bloomington, Illinois; 1923. 172 pages.

INGLIS, A. *Principles of Secondary Education*, Chapter XII. Houghton Mifflin Company, Boston; 1918.

IRION, T. W. H. "The Comprehensive Difficulties of Ninth-Grade Students in the Study of Literature." *Teachers College Contributions to Education*, No. 189. Columbia University, New York.

KIRBY, T. J. "A Grammar Test." *School and Society*, Vol. XI (June 12, 1920), pages 714-719.

KLAPPER, PAUL. "Some Observations Concerning the Measuring of Ability in Composition." *Contributions to Education*, Vol. I, Chapter VI. World Book Company, Yonkers-on-Hudson, New York; 1924.

LEWIS, E. E. *Scales for Measuring Special Types of English Composition*. World Book Company, Yonkers-on-Hudson, New York; 1921. 142 pages.

MC CALL, W. A. *How to Measure in Education*. The Macmillan Company, New York; 1921.

MCCASLIN, D. "The English Department Speaks Up at Faculty Meeting." *The English Journal*, Vol. XIV (February, 1925), pages 107-115.

MILES, D. H. "Significance of Reading in High School." *Contributions to Education*, Vol. I, Chapter XXV. World Book Company, Yonkers-on-Hudson, New York; 1924.

MONROE, W. S. "Report of Division of Educational Tests for 1919-1920." *University of Illinois Bulletin*, Vol. XVIII, No. 21; *Bureau of Educational Research Bulletin*, No. 5 (January 24, 1921), 64 pages.

— "Written Examinations versus Standardized Tests." *School Review*, Vol. XXXII (April, 1924), pages 253-265.

MORRISON, J. C. "The Use of Standard Tests and Scales in the Plattsburg High School." *University of the State of New York Bulletin*, No. 784 (July 15, 1923); 45 pages.

MORTON, R. L. "The Reliability of Measurements in Spelling." *Journal of Educational Method*, Vol. III (April, 1924), pages 321-328.

"Report of the National Committee on Reading." *The Twenty-fourth Yearbook of the National Society for the Study of Education*, Part I. Public School Publishing Company, Bloomington, Illinois; 1925. 309 pages.

RUCH, G. M. *Improvement of the Written Examination*, Chapters II and IV. Scott, Foresman & Co., Chicago; 1925.

"Sixteen Spelling Scales Standardized in Sentences for Secondary Schools." *Teachers College Bulletin*, Twelfth Series, No. 19 (May 21, 1921); 55 pages. Columbia University, New York.

STODDARD, GEORGE D. "Iowa Placement Examinations." *University of Iowa Studies in Education*, Vol. III, No. 2 (August 15, 1925); 103 pages.

THORNDIKE, E. L. "The Vocabularies of School Pupils." *Contributions to Education*, Vol. I, Chapter VII. World Book Company, Yonkers-on-Hudson, New York; 1924.

VAN WAGENEN, M. J. "The Van Wagenen Reading Scales in History, General Science, and English Literature." *Journal of Educational Research*, Vol. III (April, 1921), pages 314-316.

WILLING, M. H. "Individual Diagnosis in Written Composition." *Journal of Educational Research*, Vol. XIII, No. 2 (February, 1926), pages 77-89.

WILSON, G. M. "Language Error Tests." *Journal of Educational Psychology*, Vol. XIII (September, 1922), pages 341-349; (October, 1922), pages 430-437.

WITTY, P. A. "Treatment of Poor Spellers." *Journal of Educational Research*, Vol. XIII, No. 1 (January, 1926), pages 39-44.

CHAPTER SEVEN

SCIENCE

Introduction. Accurate measurement in the high school sciences must wait upon the development of more valid tests than those now available, and these in turn cannot appear until there is more agreement on the content of the various subjects and upon the organization of the instructional units in each subject. However, several of the tests described in this chapter will prove very useful to the science teacher, and are as valid and reliable as most standard tests in high school subjects. The extent of the field as yet inadequately treated from the standpoint of measurement in high school science has been well summarized by Glenn and Heck.¹ The statements which follow might well apply to all branches of science:

The chief purpose of any test or scale should be to improve the teaching process. If reliable tests or scales for high school physics were available, it is probable that a considerable amount of valuable evidence could be rapidly accumulated which would enable us to :

- (1) Evaluate the aims of instruction in terms of some of the measured results obtained.
- (2) Decide what items of subject matter are not learned thoroughly enough to be worthy of a place in the course.
- (3) Make a careful study of the grade placement of physics as a course or to decide which items of subject matter can be taught to the best advantage in general science or in physics.
- (4) Study the most advisable sequence of topics as represented by the recent high school textbooks.
- (5) Select drill material to make possible more complete learning of minimum essentials of the subject.
- (6) Show disinterested educators some of the contributions that physics makes in general secondary education.

¹ Glenn, Earl R., and Heck, Arch O., "Preliminary Studies of Achievement in Physics in Large City High Schools." *Contributions to Education*, Vol. I, Chapter XXX (World Book Company, 1924).

- (7) Determine the usefulness of mental tests in predicting probable school success in high school physics.
- (8) Test one teaching method against another, such as the project method versus logically arranged subject matter, or the textbook method versus a method which uses a combination of demonstration and individual laboratory experiments.
- (9) Make studies of different methods of supervising the individual study of pupils, such as the supervised study in the classroom versus the study hall, or supervised study in the classroom versus the home preparation of lessons.
- (10) Determine when to release superior students for optional work of an advanced character after certain legitimate standards have been achieved.
- (11) Determine the extent to which a student profits by classroom instruction with respect to details of the course.
- (12) Compare the achievement of classes in the same school and in different schools.
- (13) Make studies concerning the relation of the size of the class to the efficiency of instruction.
- (14) Be more intelligent in deciding whether a pupil should pass or fail in physics.
- (15) Compare the achievement of boys and girls.
- (16) Determine whether prospective college freshmen have profited by the high school course.
- (17) Classify college freshmen upon the basis of ability in the subject.
- (18) Show college professors what may reasonably be expected of good high school instruction in physics.

I. GENERAL SCIENCE

Ruch-Popenoe General Science Test

Description of the test. This test is designed to measure in one class-hour a wide sampling of the material usually offered in first-year general science. Two forms are available, each consisting of two parts. Part 1 contains 50 items covering simple information in botany, chemistry, physics, zoölogy, astronomy, physiography, geology, and physiology.

Questions are of the 7-response type. Part 2 consists of 20 diagrams and drawings, and 80 questions of the completion type. Thus the linguistic factor is reduced to a minimum, and the scoring is highly objective. The following table shows the composition of the material in the test.

TABLE 22

1. Biological science (botany, physiology, etc.)	30%
2. Chemistry	12%
3. Physics and mechanical applications	38%
4. Agriculture, astronomy, geology, etc.	20%

Administration and scoring. Administration of the Ruch-Popenoe General Science Test presents no difficulties. The actual working time is 15 minutes for Part 1 and 25 minutes for Part 2. The authors of the test found that about 90 per cent of ninth-grade pupils can attempt every item. Hence it is essentially a power test. Scoring is completely objective in Part 1, the score being the number right. Very little variation is allowed in the correct responses in Part 2, the score being one half the number right. The total possible score is 90.

Interpretation and utilization of results. The authors of the tests have indicated a number of ways in which the tests may prove of value:¹

- (1) In the assignment of school marks; by furnishing a more reliable and objective basis than that of the judgment of the teacher alone or by the use of the traditional methods of subject examinations, which have been repeatedly shown to be very unreliable.
- (2) In the determination of promotions and failures.
- (3) In the classification of pupils into sections for the purpose of differentiating rates of progress or for the enrichment of the curriculum of the abler students. In this connection it is strongly urged that a good intelligence test be given as well.

¹ Ruch, G. M., and Popenoe, H. F., "The Measurement of Ability in General Science." *School Science and Mathematics* (June, 1923), pages 545-551.

Both educational achievement and general mental ability should be utilized in sectioning classes.

(4) In comparing results obtained by the teacher of a particular class with the accomplishment of similar classes in the same school or in other localities.

Specific aids in using test results are given in the Manual of Directions.

Validity and reliability. The selection of items in the Ruch-Popenoe General Science Test was based principally on the content of common textbooks in general science. The material for Part 1 was obtained through a careful analysis of twenty-three textbooks and laboratory manuals. Only items occurring in at least half of these books were incorporated in the preliminary test, which was further refined by try-out on several hundred pupils. In selecting diagrams for Part 2 the same method of validation was employed, but was modified somewhat in accordance with the ratings of experienced teachers in general science. Part 1 provides an objective measure of the more formal content and abstract principles in general science; Part 2 parallels practical situations appearing in laboratory work. The latter measures abilities which are difficult to arrive at through ordinary teachers' examinations.

Data on the reliability of the test are given in Table 23.

TABLE 23

DATA ON THE RELIABILITY OF THE RUCH-POOPENOE
GENERAL SCIENCE TEST

r	N	S.D. ₁	S.D. ₂	P.E. _{score}	P.E. _{g.1}	P.E. _{score} S.D.	P.E. _{g.1} S.D.	NATURE OF GROUP
.83	135	14.3	—	4.2	3.6	.3	.25	High School Classes (in 4 States)
.86	44	6.16	7.80	1.75	1.62	.25	.23	High School Class

Norms. Percentile norms and *T*-scores based on about 1000 cases are given below.

TABLE 24

LATEST NORMS FOR THE RUCH-POOPENOE GENERAL SCIENCE TEST
(REVISED JANUARY 1, 1926)

PERCENTILE	MID-YEAR	END-YEAR	T-SCORES	
			RAW	T
90	42.7	53.7	80	85
80	37.3	47.5	75	81
75 (Upper quartile)	35.5	45.3	70	77
70	34.0	43.4	65	73
60	31.3	39.3	60	69
50 (Median)	28.1	35.7	55	65
40	26.9	32.3	50	60
30	24.7	29.4	45	56
25 (Lower quartile)	23.6	27.9	40	52
20	22.5	26.3	35	48
10	18.8	22.9	30	44
			25	40
N	654	1036	20	36
			15	32
			10	28
			5	24
			0	20
			N	1036

Dvorak General Science Tests

Description of the tests. These tests are published in three forms, two of which are equivalent. Each test consists of 60 5-response statements. Pupils are allowed 20 minutes or more for completing a test. Each pupil is to be permitted all the time he needs. Form R-1 is the easiest and can be used for diagnostic purposes early in the year. Forms S-2 and T-2 are equivalent and designed to measure the

achievement of a full year of general science. Items were drawn from an analysis of eighteen textbooks. These were supplemented by multiple-response statements submitted by S. R. Powers and Earl R. Glenn, and by items from the Ruch-Popenoe General Science Test. On the basis of the testing of 10,000 pupils the P.E. values of these statements were determined. Material was so arranged in the final form that two items appear at each .1 P.E. deviation from a zero-point arbitrarily fixed 8 P.E.'s below the median performance of ninth-grade pupils who had studied general science one year. Scores are weighted in accordance with these P.E. values. The probable error of estimate (.6745 S.D. $\sqrt{1-r^2}$) is about 2 scale points. Percentile norms based on 1700 ninth-grade pupils are given in the Manual of Directions.

Van Wagenen Reading Scales — General Science

These scales are available in two forms for general science. For a description of the Van Wagenen Scales, see Chapter VI. In Scale A, the 15 paragraphs are noticeably weighted toward biology and botany; in Scale B, somewhat toward chemistry and physics. Tentative percentile standards are given for eighth grade and for each year of high school.

II. BIOLOGY

Michigan Botany Test

Teachers desiring a test of botany will find the Michigan Botany Test (constructed by O. W. Laidlaw and Clifford Woody) suitable. It possesses a reliability of .87, and in 25 minutes covers a representative sample of a year course in botany. Quartile standards are given in the Manual of Directions.

Ruch-Cossmann Biology Test

Description of the test. The Ruch-Cossmann Biology Test is designed to measure accomplishment in general biology after one semester or one year of high school biology. There are five parts, as follows:

- Test 1. General biological information (40 items)
- Test 2. Incomplete statements (18 items)
- Test 3. Identification of structures from drawings (15 items)
- Test 4. Laws of Mendelian inheritance, 4 items (8 score units)
- Test 5. Completion exercises (35 items)

These items are drawn from both botany and zoölogy. Two equivalent forms, A and B, are published, each occupying one class-hour. It is recommended that the average score from both forms be used where time permits.

Administration and scoring. Complete directions for administration and scoring are given in the Manual of Directions. The time limits will be found ample. In the completion tests an unusual, but correct, word is sometimes inserted. In such cases the response is counted correct, but the statements are so worded that these occurrences will be rare. In all other respects the scoring is completely objective.

Interpretation and utilization of results. The general purposes a biology test can be made to serve closely parallel those of a general science test, as outlined on page 136. In grading and subsequent promotion or failure, sectioning of classes, and diagnosis of pupil-difficulties the test will render good service. A pupil scoring exceptionally low should be given the alternate form before judgment is passed. The test is useful too in sectioning college classes in biology on the

basis of high school performance. With the latter purpose in view the test has been made of sufficient difficulty to differentiate among college freshmen.

Validity and reliability of the test. The validation of the Ruch-Cossmann Biology Test furnishes an excellent example of one type of test validation. Items were assembled from the examination papers of 126 teachers in 23 states, selected by the state superintendents as representative of the best biology teachers in the various states. Three hundred constantly recurring questions of the two thousand received were then rated by leading teachers and authorities (77 persons). A final selection for inclusion in the test was then made from items rated as satisfactory and representative. Hence it is likely that every item in the test possesses intrinsic worth.¹ The coefficients of reliability obtained on Oregon and Iowa pupils range from .80 to .90. The probable error of a score is about 3 score units. These figures apply when only one form is used. For further data on the reliability of the Ruch-Cossmann Biology Test, see Table 25.

TABLE 25

<i>r</i>	<i>N</i>	S.D.-1	S.D.-2	P.E. score	P.E. ∞ .1	P.E. score S.D.	P.E. ∞ .1 S.D.	NATURE OF GROUP
.87	12	10.7	10.4	2.6	2.4	.25	.23	Grade X pupils
.82	18	11.2	13.4	3.5	3.2	.28	.26	Grade X pupils
.82	30	10.3	9.2	2.8	2.8	.29	.26	Grade X pupils

Norms. Percentile norms for mid-year and end-year performance for the Ruch-Cossmann Biology Test are given in Table 26. It should be borne in mind that they afford a valid means of comparison only when the ranges of talent of

¹ For a fuller account of the validation of this test, see *Journal of Educational Psychology*, Vol. XV (May, 1924), pages 285-296.

the groups compared (as measured by mental tests) are approximately equal.

TABLE 26

LATEST NORMS FOR THE RUCH-COSSMANN BIOLOGY TEST
(REVISED JANUARY 1, 1926)

PERCENTILE	MID-YEAR	END-YEAR
90	48.1	63.7
80	41.3	55.5
75 (Upper quartile)	37.3	53.0
70	34.1	49.8
60	27.2	44.2
50 (Median)	24.5	39.6
40	21.6	36.1
30	18.3	32.6
25 (Lower quartile)	16.6	30.2
20	14.7	28.2
10	10.4	23.1
N	178	753

Information Exercises in Biology (Cooprider)

This test is an outgrowth of several years' experimentation in the selection of significant questions in general biology. It is published in a single form and contains six sets of material or "exercises"; viz., Completion, Recognition, Information (5-response type), Best Reason (4-response type), Classification, and Logical Selection. The 94 items may be classified as follows:

Animal Biology	38%
Plant Biology	27%
Human Biology	18%
General Biology	17%

No time limit is set, but the majority finish in 25 to 30 minutes. Forty minutes is the maximum time allowed. Items

are not scaled and are not arranged in order of difficulty. Administration is simple; scoring is rapid and objective. Mean scores of students (total $N = 577$) having had no biology, one-half year of biology, and one year of biology are given in the Manual.

Cooprider found the reliability of the test (through Brown's formula) to be .92.¹ The correlation of test scores with teachers' estimates of biological knowledge was .70. This test should prove useful as a check upon pupil and class progress, and as an aid in classifying and promoting pupils. Although the designations of the various exercises suggest measures of mental ability, the test is valid only as an information test.

Remedial procedures in general science and biology. Pupils who are deficient in the factual portion of general science and biology should be investigated systematically. The first information needed is a measure of mental ability. This may be supplemented by the results of the Van Wagenen Reading Scales in General Science. Pupils who stand low in both these measurements cannot keep up to the standard of the class; they may be expected to remain deficient in understanding scientific vocabulary and in abstract principles and relations. Neither test gives a measure of the pupil's capacity for neat and accurate laboratory work, or for the capacity to memorize masses of data.

The reasons for poor work on the part of students mentally capable of doing the work can usually be arrived at (a) by inquiring into the pupil's work in other subjects; (b) by observation of the pupil's interests, deficiencies, and difficulties. General failure is a problem for the principal or superintendent; failure in science alone may be a reflection upon

¹ Such a coefficient cannot, of course, be interpreted without the standard deviations being known. For small classes the reliability coefficient would probably fall below the .92 given.

the type of motivation employed. F. D. Curtis¹ has recently outlined a scheme of systematic projects which in the course of several years' experimentation with about two thousand pupils has proved valuable in motivation and instruction. His plan involves:

- (a) Individual choice of an outdoor project to be selected by concurrent agreement of pupil and teacher early in the course.
- (b) Geographical delimitation of the district.
- (c) Careful preparation for all projects.
- (d) Formulation by each pupil of a complete working plan.
- (e) Inculcation of a spirit of good workmanship and of a scientific attitude.
- (f) Judgment of work according to its accuracy, difficulty, and completeness rather than on "showy" aspects.
- (g) Exhibition of work to all interested, especially parents, with provision for permanent exhibit of best projects.

Projects of this type when begun in the course in general science may be easily extended into a subsequent course in biology. They offer an excellent means for taking into account individual differences in capacity.

III. CHEMISTRY

Introduction. The few thoroughgoing studies that have been made of the teaching of high school chemistry indicate an urgent need for curriculum organization. It is not clear what is expected of chemistry teachers, and the textbooks and manuals upon which courses are usually based indicate wide divergence. It is small wonder that there are really no good tests in chemistry as yet, although some promising

¹ Curtis, F. D., "Systematic Project Work in General Science and Biology," *School Science and Mathematics*, Vol. XXIV, No. 9 (December, 1924), pages 968-974.

beginnings have been made. The work of E. R. Glenn and L. E. Welton at The Lincoln School of Teachers College, New York, represents an attempt to survey in a comprehensive manner chemistry content. The tests they have developed are available in experimental form for schools wishing to coöperate.

Textbooks in chemistry are not a safe guide to either content or method. The excellence of preparation for college chemistry cannot be judged by the amount of ground covered or by the grades received in high school chemistry. For evaluation according to the criterion of success in college chemistry the Iowa Placement Examinations will be found useful.

The aims of high school instruction in chemistry which are commonly advanced are:

- (1) College preparation.
- (2) Knowledge of principles, facts, applications of chemistry.
- (3) Relation of chemistry to daily life.
- (4) Training to think (in chemistry).

Powers's study¹ indicates that such ends were not actually attained in the cases he investigated (about two thousand in number), and that the textbooks are not necessarily designed in such a way as to offer a hope of accomplishing what they set out to do. Thus chemistry is another subject where the teacher should become skillful in building up his own objective examinations. A test cannot be properly standardized and norms cannot have much meaning until the scope and the aims of a subject are fairly well defined.

¹ Powers, S. R., "A Diagnostic Study of the Subject Matter of High School Chemistry." *Teachers College Contributions to Education*, No. 149 (Columbia University, New York, 1924).

Powers General Chemistry Test

Description of the test. Part I consists of 30 items covering range of information in biography, chemical processes, and terminology. Part II, containing 37 items, tests the pupil's ability with respect to formulas and equations, chemical names of substances, and simple calculations. The test appears in two equivalent forms, A and B.

Administration and scoring. The working time for each form is 35 minutes. It is essentially a power test, since increasing the time allowance has little effect upon the scores obtained. Scoring may be done rapidly and objectively by means of the printed key.

Interpretation and utilization of results. The author of the test lists several ways in which the tests prove helpful to chemistry teachers:

- (1) In the assignment of school marks.
- (2) In determining promotions and failures.
- (3) In comparing classes and schools.
- (4) In predicting success with college entrance examinations and, to some extent, in college work in chemistry.

Validity and reliability. The items were selected originally from common high school textbooks in chemistry. Experimental work over a period of four years reduced the number of items from 350 to 134 in the two forms. Items are arranged in ascending difficulty in accordance with an item-count based on a minimum of 418 cases. Sixty schools coöperated. Table 27 gives reliability data (adapted from values submitted to the authors by Professor Powers).

TABLE 27

DATA ON RELIABILITY OF POWERS GENERAL CHEMISTRY TEST
(Figures are for equivalent forms of 30 items each)

<i>r</i>	<i>N</i>	S.D. ₁	P.E.-score	P.E. _{∞,1}	P.E.-score S.D.	P.E. _{∞,1} S.D.	NATURE OF GROUP
.84	101	6.1	2.04	1.89	.33	.31	11th and 12th grades
.74	68	6.0	2.06	1.82	.34	.30	11th and 12th grades
.79	56	9.5	2.94	2.62	.31	.27	11th and 12th grades

Norms. The following are medians for juniors and seniors in various high schools, the standard median being 88.

NUMBER OF CASES	MEDIAN
27	87.7
35	91.5
82	84.3
46	96.0
39	91.0
113	85.5
72	90.8
63	76.1
43	92.7
136	91.1
10	86.5

Composite percentile norms are given in the Manual of Directions.

*Iowa Placement Examinations*¹

Two chemistry tests are available in the Iowa Placement series: Chemistry Aptitude, CA-1, Revised (2 forms), and Chemistry Training, CT-1, Revised (2 forms). The composition of these examinations is as follows:

¹ See Chapter XI for a complete description of the validity, reliability, and utility of these examinations.

Chemistry Aptitude, CA-1, Revised

Part 1 (20 items) is devoted to the simple arithmetic of chemistry. It involves problems which in the experience of a teacher usually give difficulty to beginning students. In Part 2 (30 items) selections are made from textbooks in college chemistry and the student is asked precise questions as to content and relations in the paragraph. This is intended to measure the student's ability for exactness of reading, combined with ability to resist generalizations beyond the data. Part 3 (15 items) is an adaptation of the method employed in the Iowa Comprehension Test (see Chapter XI) to the measurement of chemistry reading comprehension. The material in this part is fairly difficult, and the student is expected to show that he has a grasp of the ideas contained in it as well as of the factual elements. Part 4 (60 items) attempts to measure interest in chemistry through the accuracy of the student's common information in chemistry. The assumption is that students who have particular fitness and liking for the subject will tend to read the semi-popular journals connected with it and will in various ways acquire a considerable body of chemical information. The total working time of CA-1, Revised, is 44 minutes.

Chemistry Training, CT-1, Revised

Part 1 measures knowledge of fundamentals of chemical processes. It consists of 45 questions of the true-false type, representing a sample of the material from textbooks commonly used in college. Part 2 consists of 45 questions of the recall type, covering valencies, formulas, names of compounds, and the completion and balancing of equations. Part 3 consists of 50 questions of the true-false type, emphasizing manufacturing processes and the applications of

chemistry. Part 4 consists of 12 fundamental chemistry problems, each bringing out different phases of computation. The mechanics of arithmetic are reduced to a minimum, so that a student who is familiar with the relations involved has no difficulty in doing the questions in the allotted time. The total working time is 43 minutes.

OTHER TESTS IN CHEMISTRY

Chemistry Tests Gamma and Epsilon were devised by S. G. Rich to measure achievement in chemistry. Each form consists of 25 4-response items in general chemistry, requiring 25 minutes' working time. The material was drawn from common textbooks, and in some cases from college entrance examinations. An attempt was made also to have the test conform to social aims. Norms in half-semester increments are given up to four semesters, both high school and college, in the Manual of Directions. They are given in *T*-scores, a conversion table being provided in the Manual. Reliability coefficients submitted by the author run from .55 to .70. The probable error of a score is about 4 *T*-score units. Since the difference of attainment between one semester and two semesters of high school chemistry is given as 8.8 *T*-score units, individuals cannot be considered accurately measured by this test. The 25 items of the Rich Chemistry Test are designed to measure not only individual and group attainment in general, but such phases as ability to think, information, ability to solve numerical problems, and "habits and knowledge acquired from work in the laboratory." The test is not of sufficient length to accomplish these ends.

The experimental work of E. R. Glenn and Louis E. Welton in constructing objective examinations in chemistry is of interest. Thirty-six tests are published in one booklet,

designed to cover every topic in high school chemistry. The tests are not on the market, but teachers wishing to coöperate in the general experiment may communicate with Dr. E. R. Glenn, The Lincoln School of Teachers College, Columbia University.

IV. PHYSICS¹

Iowa Physics Tests

Description of the tests. Three tests (they are in fact scales), devised by H. L. Camp, measure the following branches:

Series A, Forms 1 and 2, Mechanics.

Series B, Forms 1 and 2, Heat.

Series C, Forms 1 and 2, Electricity and Magnetism.

Each form is a 2-page folder containing 11 or 12 questions. The working time is 45 minutes. The questions are of the recall type: the chance element is eliminated. The purpose of the test as given by the author is to measure

- (a) Knowledge of the fundamental principles of physics;
- (b) Ability to put this knowledge to use in solving problems one meets in ordinary life.

Interpretation and utilization of results. These tests provide an objective measurement of a portion of physics, and can be used in a comparative way. Norms (medians and interquartile ranges) are provided in the Manual of Directions. The norms are not sufficiently extensive to throw much light on individual pupil-progress.

Validity of the tests. Test exercises were based on 102 principles found by Daniel Starch to be common to five high school physics textbooks. Thirty-five hundred pupils in 129 Iowa high schools worked on the preliminary tests. In the final forms the items have been arranged in order of

¹ See Chapter XI for the Columbia Research Bureau Physics Test.

difficulty and weighted accordingly in the scoring. This method of weighting decreases the value of the tests, for it is based upon the fallacy that the least-known facts and problems in physics are the most important.

Hughes Physics Scales

The Hughes Physics Scales are four in number, two designed to measure information and two to measure thought. The Information Scale requires 20 minutes, the Thought Scale, 37 minutes. The Information items were selected on the basis of returns on a questionnaire sent to physics teachers, supplemented by textbook analyses. Thought questions involve reasoning and a minimum of routine computation. However, there is not much distinction between the Information Scale and the Thought Scale. Thus the question, "What is the common name of the process by which the sun transmits heat to the earth?" is in a "Thought" Scale, while the question, "What reading on a Fahrenheit thermometer corresponds to a -40° reading on the Centigrade thermometer?" is in an "Information" Scale. The Hughes Physics Scales were formed in accordance with a rigorous P.E. scheme. Each scale yields a corrected score which gives the difficulty of the item each pupil can do with a correctness of 50 per cent. The utility of these scales is similar to that of the Iowa Physics Tests. They are the result of more recent experimentation than the latter.

Iowa Placement Examinations¹

Two physics tests are available in the Iowa Placement series: Physics Aptitude, PA-1, Revised (2 forms), and Physics Training, PT-1, Revised (2 forms). The composition of these examinations is as follows:

¹ See Chapter XI for a complete description of the validity, reliability, and utility of these examinations.

Physics Aptitude, PA-1, Revised

Part 1 (25 items) measures the simple arithmetic of physics. It is similar to the corresponding part in Chemistry Aptitude. Part 2 (30 items) is similar to the corresponding part in chemistry. It consists of paragraphs calling for precise statements which show the student's knowledge of the material in the paragraph. Part 3 (20 items) combines a brief number series with a test of logic similar to that found in Mathematics Aptitude. The relative value of this part in comparison with the reading comprehension test found in Chemistry Aptitude, CA-1, will be determined. Part 4 (50 items) is an interest test. It assumes that one's interest will be to a certain extent measured by one's fund of common knowledge of the subject and that this interest will in turn be a factor in subsequent performances in physics. The assumptions here are analogous to those in Part 4 of Chemistry Aptitude. The total working time is 45 minutes.

Physics Training, PT-1, Revised

Part 1 consists of 40 statements of the completion type on fundamental information and principles in physics. Part 2 (50 items) consists of material of the same general nature, arranged as true-false items. Parts 1 and 2 were drawn through analysis of common college and high school textbooks in elementary physics and cover the same material. The two types of objective examinations are used because certain questions lend themselves more readily to one type than to another. Part 3 (20 items) calls for the completion of physical equations and knowledge of certain important applications and principles in physics. Part 4 is a series of 15 problems in physics, each touching upon an important phase. Computation is reduced to a minimum. The total working time is 43 minutes.

Thurstone Vocational Guidance Tests — Physics¹

The Physics Test of the Thurstone Vocational Guidance series consists of 25 problems and requires 30 minutes' working time. Items were selected which "appeal to the engineering interests of the candidate. It is not sufficient that the student be able to repeat the principles of physics; he must also be able to apply these principles on his own initiative to problems of engineering interest even when the problems are so stated that the principle involved is not immediately apparent. The student who has memorized the subject matter of the courses in physics without realizing its significance will be handicapped in taking this examination. This is as it should be, because the subject matter is in that case not functioning."² This test correlated .34 with average freshman scholarship for engineering students, which is about the same as the relationship between high school physics marks and college freshmen scholarship.

Remedial procedures in chemistry and physics. Diagnosis of pupil difficulties in chemistry should be carried on by the teacher through the development of objective examinations covering in some detail the various phases of the year's work. The booklet, *New-Type Chemistry Tests*, by E. R. Glenn and L. E. Welton, will prove serviceable as source material. A test of mental ability will be necessary where failure is chronic. The standard tests available will give a comparative measure at the end of a year of chemistry, and will give in addition a fair indication of the likelihood of success in college chemistry. The results from the Iowa Placement Examinations in chemistry have indicated that the common causes of failure in first-year chemistry are deficiencies in mental ability, reading comprehension,

¹ See Chapter X for a description of these tests.

² Thurstone, L. L., *Manual of Directions for the Thurstone Vocational Guidance Tests* (World Book Company).

and arithmetical ability. Causes of failure in the second course are: (a) inadequacy of previous preparation (it has been found that previous instruction in high school chemistry is no guarantee that the student can go on to more advanced chemistry); (b) lack of interest and application.

The teacher of physics will find the scales now on the market inadequate in diagnosing pupil difficulties. What is needed is a complete series of tests, perhaps of an inventory nature, validated on social grounds. The work of E. R. Glenn¹ and E. S. Obourn in preparing and trying out 52 physics tests in 1922-24 is of considerable importance to physics teachers. There are two forms, each comprising a printed booklet of 26 tests, covering mechanics, heat, electricity, sound, and light. Such tests provide rich source material in the preparation of objective examinations for instructional and measurement purposes. Iowa Placement Examinations, Physics Aptitude and Physics Training, given near the end of the senior year will enable comparisons with the general college freshman level of aptitude and training and will indicate the likelihood of subsequent success in the subject. The Thurstone Vocational Guidance Test in Physics (see Chapter X) will render similar service, particularly for engineering students.

¹ Glenn, Earl R., and Heck, Arch O., "Preliminary Studies of Achievement in Physics in Large City High Schools." *Contributions to Education*, Vol. I (World Book Company, 1924).

Test Materials

Dvorak General Science Tests. By AUGUST DVORAK. Forms R-1, S-2, and T-2, each 50 cents per package of 25, including Manual of Directions. Specimen set, 20 cents. Public School Publishing Company, Bloomington, Illinois.

Ruch-Popenoe General Science Test. By G. M. RUCH and H. F. POPENOE. Forms A and B, each \$1.30 per package of 25, including Manual of Directions, Key, Percentile Graph, and Class Record. Specimen set, 20 cents. World Book Company, Yonkers-on-Hudson, New York.

Van Wagenen Reading Scales—General Science. By M. J. VAN WAGENEN. Forms A and B, each \$3.00 per 100. Specimen set, 20 cents. Public School Publishing Company, Bloomington, Illinois.

Cooprider Information Exercises in Biology. By J. L. COOPRIDER. 50 cents per package of 25, including Manual of Directions. Specimen set, 10 cents. Public School Publishing Company, Bloomington, Illinois.

Ruch-Cossmann Biology Test. By G. M. RUCH and L. H. COSSMANN. Forms A and B, each \$1.30 per package of 25, including Manual of Directions, Key, and Class Record. Specimen set, 20 cents. World Book Company, Yonkers-on-Hudson, New York.

Powers General Chemistry Test. By S. R. POWERS. Forms A and B, each \$1.10 per package of 25, including Manual of Directions, Key, and Class Record. Specimen set, 20 cents. World Book Company, Yonkers-on-Hudson, New York.

Rich Chemistry Test for High Schools. By S. G. RICH. Forms Gamma and Epsilon, each \$1.00 per package of 25, including Directions and Class Records. Specimen set, 20 cents. Public School Publishing Company, Bloomington, Illinois.

Hughes Physics Scales. By J. M. HUGHES. Information R, Division I; Information S, Division II; Thought R, Division I; Thought S, Division II; each Division 50 cents per package of 25, including Class Record. Specimen set, 15 cents. Public School Publishing Company, Bloomington, Illinois.

Iowa Physics Tests. By HAROLD L. CAMP. Series A (Mechanics), B (Heat), and C (Electricity and Magnetism), Forms 1 and 2 for each series, each form 50 cents per package of 25, including Manual of Directions and Class Record. Specimen set, 15 cents. Public School Publishing Company, Bloomington, Illinois.

Michigan Botany Test. By O. W. LAIDLAW and CLIFFORD WOODY. \$1.00 per package of 25, including Direction Sheet. Specimen set, 15 cents. Public School Publishing Company, Bloomington, Illinois.

Iowa Placement Examinations.

Chemistry Aptitude, CA-1, Revised, Forms A and B
 Chemistry Training, CT-1, Revised, Forms A and B
 Physics Aptitude, PA-1, Revised, Forms A and B
 Physics Training, PT-1, Revised, Forms A and B

Each form \$3.50 per 100, with Manual and Key. Bureau of Educational Research and Service, University of Iowa, Iowa City, Iowa.

Thurstone Vocational Guidance Tests. By L. L. THURSTONE. Physics Test, \$1.00 per package of 25, with Key and Record Sheet. World Book Company, Yonkers-on-Hudson, New York.

References

BARBER, FRED D. "The Reorganization of High School Science." *School Science and Mathematics*, Vol. XXIII (1923), pages 247-262.

CALDWELL, OTIS W., and Committee. *Reorganization of Science in Secondary Schools*. Bulletin No. 26 (1920); 62 pages. Department of the Interior, Bureau of Education, Washington, D. C.

CAMP, H. L. "Scales for Measuring Results of Physics Teaching." *Journal of Educational Research*, Vol. V (May, 1922), pages 400-405.

— "An Evaluation of Standard Tests and Suggested Uses in Improving Physics Teaching." *School Science and Mathematics*, Vol. XXIII (1923), pages 441-446.

COOPRIDER, J. L. "Information Exercises in Biology." *School Science and Mathematics*, Vol. XXV, No. 8 (November, 1925), pages 807-813.

CORNOG, JACOB, and STODDARD, GEORGE D. "Predicting Performance in Chemistry." *Journal of Chemical Education*, Vol. II (August, 1925), pages 702-708.

DOWNING, E. R. "The Revised Norms for the Range of Information Test in Science." *School Science and Mathematics*, Vol. XXVI (February, 1926), pages 142-146.

GERRY, H. L. "Types of Tests Desirable for Chemistry and the Present Status of Their Development." *School Science and Mathematics*, Vol. XXV (September, 1925), pages 918-922.

GLENN, EARL R. *Bibliography of Science Teaching in Secondary Schools*, Bulletin No. 13 (1925); 161 pages. Department of the Interior, Bureau of Education, Washington, D. C.

— and HECK, ARCH O. "Preliminary Studies of Achievement in Large City High Schools." *Contributions to Education*, Vol. I, Chapter XXX. World Book Company, Yonkers-on-Hudson, New York; 1924.

MAXWELL, P. A. "Standardizing of First-Year Science Tests." *General Science Quarterly*, Vol. V (1921), pages 226-231.

MILLIKAN, R. A. "The Problem of Science Teaching in the Secondary Schools." *School Science and Mathematics*, Vol. XXV, No. 9 (December, 1925), pages 966-975.

POWERS, S. R. "Tests of Achievement in Chemistry." *Journal of Chemical Education*, Vol. I (1924), pages 139-144.

— "Achievement in High School Chemistry — An Examination of Subject Matter." *School Science and Mathematics*, Vol. XXV (1925), pages 53-62.

— "The Vocabularies of High School Science Textbooks." *Teachers College Record*, Vol. XXVI (1925), pages 368-383.

"Report of the Committee on Reorganization of the Biological Sciences, Appointed by the Cleveland Biology Teachers' Club." *School Science and Mathematics*, Vol. XXIV (1924), pages 241-246.

RICH, S. G. "Achievements of Pupils in Chemistry." *School Science and Mathematics*, Vol. XXV (1925), pages 145-149.

RIVETT, B. J. "Results with Standard Chemistry Tests." *School Science and Mathematics*, Vol. XXI (1921), pages 720-722.

RTCH, G. M. "A Range of Information Test in General Science." *General Science Quarterly*, Vol. IV (November, 1919), pages 257-262.

— "A Range of Information Test in General Science: Preliminary Data on Standards." *General Science Quarterly*, Vol. V (November, 1920), pages 15-19.

— "Tests and Measurements in High School Science." *School Science and Mathematics*, Vol. XXIII, No. 9 (December, 1923), pages 885-891.

— and COSSMANN, L. H. "Standardized Content in High School Biology." *Journal of Educational Psychology*, Vol. XV, No. 5 (May, 1924), pages 285-296.

— and POPENOE, H. F. "The Measurement of Ability in General Science." *School Science and Mathematics*, Vol. XXIII (June, 1923), pages 545-551.

STODDARD, GEORGE D. "Iowa Placement Examinations." *University of Iowa Studies in Education*, Vol. III, No. 2 (August 15, 1925); 103 pages.

TOOPS, H. A. "A General Science Test." *School Science and Mathematics*, Vol. XXV, No. 8 (November, 1925), pages 817-822.

CHAPTER EIGHT

FOREIGN LANGUAGE

I. FRENCH AND SPANISH

Introduction. Measurement in French and Spanish is in a state of flux. The test materials now available for the high school teacher are few in number and rather narrow in scope. However, great projects (chief of which is the Modern Foreign Language Study) are under way, and these investigations should produce valuable source material for the construction of standard tests. At the present writing the most significant work in language testing is that which centers about Henmon's French word count at the University of Wisconsin, the Placement Examinations developed at Columbia University and the University of Iowa, and the French, German, and Spanish tests under development as a part of the Modern Foreign Language Study. The Iowa Placement Examinations (French Training, Spanish Training, and Foreign Language Aptitude) have been widely used by colleges in evaluating the modern language work of high schools.

Henmon French Tests

Description of the tests. Four equivalent tests are published, each consisting of a vocabulary and a sentence section printed on opposite sides of the same sheet. The vocabulary section consists of 60 words drawn from 448 words drawn from 12 widely used French texts. Each word is weighted according to its P.E. value. The sentence section (12 sentences) contains only words in the list of 448, and each sentence also is given a scale value. The working time of the test is undetermined, but is usually 20 minutes.

Interpretation and utilization of results. Half-year norms up to 3 years of high school French are given in the Manual of Directions. The test is non-diagnostic and too brief to

give more than a rough measure of class-standing. Reliability data are given in Table 28.

TABLE 28

DATA ON THE RELIABILITY OF THE HENMON FRENCH TESTS
(TEST 1 vs. TEST 2)

<i>r</i>	<i>N</i>	S.D.1	S.D.2	P.E.-score	P.E.- <i>x</i> .1	P.E.-score S.D.	P.E.- <i>x</i> .1 S.D.
.61	60	51.5	60.3	23.5	18.4	.40	.33

Handschin Modern Language Tests

The Silent Reading Test A : French, consists of 12 exercises. The pupil reads each French sentence and answers in French the question contained in it; 5 minutes are allowed. Silent Reading Test B : French, consists of a French paragraph (192 words). The pupil reads the French and then is allowed 5 minutes to answer, in French or English, 10 questions based on the paragraph. The option of French or English in the responses introduces undesirable variations. The Silent Reading Tests are for first-year or second-year French. Comprehension and Grammar Test A : French (for first-year French), requires various completions and inflections in 6 easy French sentences. The working time of this test is 10 minutes. The Silent Reading Test A : Spanish, is similar to the corresponding French test. All the Handschin Modern Language Tests are non-diagnostic and too lacking in comprehensiveness to be of much real assistance to the teacher of French.

Columbia Research Bureau French Test

Description of the test. Forms A and B of the Columbia Research Bureau French Test have recently been published

by the World Book Company. Dr. Ben D. Wood is largely responsible for the careful work leading to its validation. The comprehensiveness of the test is indicated by a tabulation of its contents:

PART	NAME	NUMBER OF QUESTIONS	MINUTES
I	Vocabulary	100	25
II	Comprehension	75	20
III	Grammar	100	45

Where less than 90 minutes' testing time is available, the parts can be given separately. The reliability of the test is .96 for the 4-year range of French classes. (See also Chapter XI for the Columbia Research Bureau Tests in French, Spanish, and German.)

*Iowa Placement Examinations*¹

French Training, FT-1, Revised

Description of the examination. The French Training Examination can be given after one-half year of French in high school or college, and covers a range of talent sufficient to measure four years of high school French.

Part 1 consists of 60 French words for which the English equivalent is to be given. The words were taken systematically from the recently published French word count of V. A. C. Henmon. The words are given in ascending order of difficulty. Part 2 is a test of French grammar, consisting of 40 items. These items contain most of the points of grammar set forth by the Committee on Syllabi of the Association of Modern Language Teachers of the Central West and South. In Part 3 (40 items), which tests French idioms and tenses of verbs, the recommendations of the committee referred to above are again followed. Part 4 (20 items)

¹ See Chapter XI.

measures French reading comprehension. Three paragraphs are presented in increasing difficulty and questions are asked in English after each paragraph. There is little possibility of guessing the correct answers, and the questions are so asked that a brief response can be given in each case.

The total working time of FT-1, Revised, is 45 minutes.

Spanish Training, ST-1, Revised

Description of the examination. The Spanish Training Test of the Iowa Placement Examination series is analogous to the French Training Test. In Part 1, 50 words are listed, drawn from a word count by V. A. C. Henmon. The English equivalent is asked for in a 5-response situation. Part 2 (40 items) is a test of Spanish grammar, the examples being chosen in accordance with the recommendations of the Committee on Syllabi of the Association of Modern Language Teachers of the Central West and South. Part 3 (40 items) is a test of verb forms covering the principal usages of common verbs. Part 4 parallels the reading comprehension test of French Training, FT-1, Revised. Three paragraphs are given in ascending order of difficulty, and definite questions are asked concerning them. Answers are to be given in English. The total working time of ST-1 is 43 minutes.

Wilkins Prognosis Test in Modern Languages

Description of the test. This test is designed for the purpose of "making as manifest as possible the probable fitness of those examined for pursuing foreign language studies." Six tests comprise the following measurements:

- I. Visual-motor (seeing and writing)
- II. Aural-motor (hearing and writing)
- III. Memory
- IV. Grammar Concepts
- V. Visual-oral (seeing and speaking)
- VI. Aural-oral (hearing and speaking)

A brief auxiliary test in French and Spanish is printed on the test booklet as an aid in subsequent shifting of pupils improperly placed.

Administration and scoring. Tests V and VI are individual tests; the others are group tests. Teachers planning to give this test should order the material well in advance, for flashcards must be made out, and a certain amount of practice in giving the test is desirable. The time required is 24 minutes for all four group tests, and about 2 minutes for the two individual tests. Tests I, II, V, and VI are scored rather awkwardly and subjectively.

Interpretation and utilization of results. Wilkins states that students scoring less than 360 (60 per cent of the maximum score) are not likely to succeed in modern language work. However, it is unlikely that this test is sufficiently valid and reliable to make prediction of individual performance as definite as that. Thus Jordan¹ found that for 81 pupils the test correlated with teachers' marks in foreign language: $r = .75$, but that the correlation dropped to .49 for a different group of 108 pupils. The reliability of the prognosis test and of the teachers' marks and the range of talent of the group investigated are important variables in such predictions. The Wilkins Prognosis Test, especially in conjunction with a reliable measure of intelligence, is more dependable in properly sectioning classes for differential instruction than in predicting the success or failure of individuals in the group.

Iowa Placement Examinations

Foreign Language Aptitude, FA-1, Revised

Description of the examination. Foreign language aptitude is here measured by recourse to skills involved in the

¹ Jordan, J. N., "Prognosis in Foreign Language in Secondary Schools." *School Review*, Vol. XXXIII, No. 7 (September, 1925), pages 541-546.

manipulation of English and Esperanto. Results obtained with this examination are tabulated in Chapter XI. The content of the examination is indicated in the paragraph which follows:

Part 1 (50 items) is a measure of the elements of English grammar, with special reference to parts of speech, inflections, and the roots of common English words. Part 2 (40 items) measures the amount of transfer of training from English to an unfamiliar language. Esperanto is employed, and in most cases the student's success in recognizing the significant word in the Esperanto sentence depends on his ability to grasp the probable meaning of the whole unit of thought. In Part 3 (30 items) Esperanto is employed to measure the student's ability to comprehend and apply rules of grammar. Three rules are given which deal with the formation of words from roots and with the number system in Esperanto. Part 4 (30 items) is a measure of aptitude for translation. The student observes an English translation of material in Esperanto and is asked for English equivalents, Esperanto equivalents, and parts of speech. The total working time of FA-1, Revised, is 45 minutes.

Remedial procedures in French and Spanish. The extensive investigations in modern foreign language now being conducted are sure to throw much light on teaching practice, and may reasonably be expected to point out the best method of accomplishing definite aims in language work. In the meantime valuable remedial measures may result from better knowledge of the pupil's capacity and training deficiencies:

(a) The teacher should build up and frequently employ brief objective tests of vocabulary, grammar, and reading comprehension. Illustrative tests of this type will be found in the pamphlet, *Course of Study in French for High Schools*, by Helen M. Eddy. Vocabulary tests can be constructed from Henmon's *French Word Lists*, or from the

Ward *Minimum French Vocabulary Test Book*. The Ward list gives French-to-English translations of the 2000 commonest words, together with useful pupil-motivation devices.

(b) Only the longer standard French and Spanish tests mentioned furnish valid measures of achievement; they too are to some extent diagnostic in the grammar sections.

(c) For a suggestive summary of concrete helps in language instruction the teacher is referred to C. E. Young's booklet, "French and Spanish in the High School," *University of Iowa Extension Bulletin*, No. 93, September 1, 1923.

II. LATIN

Introduction. The older Latin tests have proved useful in affording brief comparative measures of class-attainment but have given little diagnostic information. Tests more recently developed, such as the Godsey Diagnostic Test and the Ohio State Series, may be expected to carry increasingly the weight of Latin testing. They represent, however, only a small beginning; and every Latin teacher should undertake to devise his own informal objective tests; nation-wide norms are not needed to render these tests valuable aids in discovering pupil-deficiencies and indicating the kind and extent of remedial drill. The best authority for determining the essentials of Latin instruction, and hence the valid materials for Latin testing, is the American Classical League's report, *The Classical Investigation*.¹

Henmon Latin Tests

Description of the tests. These tests parallel the Henmon French Tests in appearance and method of construction.

¹ West, Andrew F. (Chairman), *The Classical Investigation*, Part I (Princeton University Press, Princeton, New Jersey; 305 pages).

Four equivalent forms, Tests 1, 2, 3, and 4, afford a brief measure of Latin vocabulary and of Latin-to-English translation. The 50 words in the vocabulary test were drawn from 239 words common to thirteen widely used first-year Latin texts — Cæsar, Cicero, and Vergil. The 10 sentences in each test were devised from the selected vocabulary. Text X of the Henmon Latin Tests is not equivalent to the other four, but is designed for research purposes. The working time for a single test is about 20 minutes.

Interpretation and utilization of results. Norms are given in the Manual of Directions for each year of high school Latin. The vocabulary test is too easy for class-discrimination beyond the third semester of high school Latin. For individual measurement, all four tests should be given; otherwise the unreliability is great. Data on the reliability of the Henmon Latin Tests are given in Table 29.

TABLE 29

DATA ON THE RELIABILITY OF THE HENMON LATIN TESTS¹

<i>r</i>	<i>N</i>	S.D.	S.D.	P.E. score	P.E. _{g.1}	P.E. score	S.D.	P.E. _{g.1}	S.D.	NATURE OF GROUP
<i>Vocabulary</i>										
.66	47	13.6 (1)	17.2 (2)	6.1	4.9	.40	.31			High School
.68	44	17.5 (2)	14.6 (3)	6.1	5.0	.38	.31			Year I to Year
.74	44	13.6 (1)	14.6 (3)	4.8	4.2	.34	.29			IV (Pooled)
.75	43	13.6 (1)	17.9 (4)	5.3	4.7	.34	.29			
.80	41	15.0 (3)	17.3 (4)	4.9	4.4	.30	.27			
<i>Sentences</i>										
.50	47	5.6 (1)	9.0 (2)	3.5	2.5	.48	.34			High School
.71	44	9.5 (2)	9.6 (3)	3.5	3.0	.36	.31			Year I to Year
.53	44	5.4 (1)	8.5 (3)	3.2	2.4	.46	.34			IV (Pooled)
.54	43	5.7 (1)	9.4 (4)	3.5	2.6	.46	.34			

¹ The number of the test form is indicated in the S.D. column.

Ullman-Kirby Latin Comprehension Test

Description of the test. The Ullman-Kirby Test consists of ten Latin paragraphs of increasing difficulty, with each paragraph followed by three or four questions in English. Answers are to be given in English. The material is typical of Latin readings in a 4-year course. The working time of the test is 30 minutes. Scoring is somewhat subjective.

Interpretation and utilization of results. The Classical Investigation (General Report, Part I, page 194) found that the Ullman-Kirby Latin Comprehension Test as a measure of comprehension correlated very highly with the test as a measure of translation. In addition, the emphasis placed upon grasping the whole thought of a sentence or paragraph is desirable. The test can be used to mark class-progress and for comparison of group and school performances. It is useful, too, for purposes of class-sectioning. For individual diagnosis of Latin comprehension, it is recommended that additional objective measures be prepared by the teacher. Norms are given in Table 30, and data on reliability in Table 31. Additional norms are contained in the Manual of Directions.

TABLE 30

NORMS, ULLMAN-KIRBY LATIN COMPREHENSION TEST¹

FORM→	SEMESTER								SEMESTERS COMBINED	
	II		IV		VI		VIII			
FORM→	I	II	I	II	I	II	I	II	I	II
No.	123	123	101	101	109	109	90	90	423	423
Mean	11.2	10.7	17.4	18.2	21.6	23.1	24.8	24.7	18.5	18.8
S.D.	2.5	4.0	4.1	3.7	4.8	4.0	4.7	4.0	6.6	7.0

¹ From Ullman, B. L., and Kirby, T. J., "A Latin Comprehension Test." *Journal of Educational Research*, Vol. X (November, 1924), pages 308-318.

TABLE 31

DATA ON THE RELIABILITY OF THE ULLMAN-KIRBY LATIN COMPREHENSION TEST

NO. SEMESTERS LATIN	N	r	S.D. ₁	S.D. ₂	P.E-score	P.E. _{∞.1}	P.E-score S.D.	P.E. _{∞.1} S.D.
2	123	.53	2.5	4.0	1.5	1.1	.47	.34
4	101	.65	4.1	3.7	1.5	1.2	.39	.31
6	109	.57	4.8	4.0	1.9	1.5	.43	.34
8	90	.71	4.7	4.0	1.6	1.3	.37	.30
2-8 Combined	423	.85	6.6	7.0	1.8	1.6	.27	.23
2-8 (High School Classes)	43	.66	3.8	4.8	1.7	1.4	.38	.32

White Latin Test

Description of the test. The White Latin Test consists of a vocabulary of 100 words (in a 4-response situation) and the translation of 20 sentences. It is designed to cover the 4-year range of high school Latin. The working time is 35 minutes. Scoring is completely objective.

Interpretation and utilization of results. Norms (medians only) are given in the Manual of Directions. The test will give a rough measure of Latin attainment for the purpose of class-sectioning. That the two forms are not equivalent, and the test not sufficiently reliable for individual pupil measurement, is shown in Table 32.

TABLE 32

DATA ON THE RELIABILITY OF THE WHITE LATIN TEST

r	N	S.D. _A	S.D. _B	P.E-score	P.E. _{∞.1}	P.E-score S.D.	P.E. _{∞.1} S.D.
.38	67	13.4	16.7	6.3	4.9	.41	.33

Tests in Latin Vocabulary, Latin Derivatives, Latin Verb Forms, and Latin Syntax

Description of the tests. A series of four Latin tests has been developed at Ohio State University, as follows:

NAME OF TEST	AUTHOR	NO. ITEMS	WORKING TIME	NO. FORMS	STANDARDS
(1) Latin Vocabulary Test . .	P. R. Stevenson	60	15 min.	3	Medians, Gr. VII-IX
(2) Latin Derivative Test . .	P. R. Stevenson and W. W. Coxe	60	15 min.	3	Not given Medians
(3) Latin Verb Forms	C. Tyler and S. L. Pressey	32	20 min.	1	2-8 semesters
(4) Latin Syntax	L. W. Pressey	32	20 min.	1	2-8 semesters

These tests can be used to discover pupil weaknesses and to indicate the need for remedial drill. The Classical Investigation reports various findings of interest on the basis of the Pressey tests mentioned above.

Godsey Latin Composition Test

Description of the test. The Godsey Latin Composition Test was developed in connection with the Classical Investigation. Two forms have been published, each requiring 35 minutes' working time. The test can be given after one year of Latin and is designed to test the whole range of high school Latin. Latin words in the test were based on Henmon's list of 239 words common to beginning Latin readings. The test covers sentences and rules of syntax. Scoring is completely objective.

Interpretation and utilization of results. Median scores based on 20,000 cases are given in the Manual of Directions for both sentences and rules, but uses of the test are intended

rather to be diagnostic. The author states that the test "is diagnostic mainly with reference to the general tendencies manifested by a class as a whole. It has less significance when applied to the errors of an individual pupil." The test is so constructed that only one type of error is possible in each incorrect answer. Thus the teacher can make a tabulation of the commonest errors found in his class. These indicate the weak points of the class, and drill work can be adjusted accordingly.

Orleans-Solomon Latin Prognosis Test

Description of the test. The Orleans-Solomon Latin Prognosis Test presents to the high school pupil about to begin the study of Latin a series of tasks involving simple learning in Latin. The total working time is 51 minutes. The test is described in detail in the Manual of Directions.

Interpretation and utilization of results. This test may be expected to predict high school grades in Latin to the extent represented by a correlation of .75. However, great care should be exercised in eliminating pupils from Latin classes on the basis of test results alone. General intelligence, interest, and the probable future vocation of the pupil should be taken into account.

COMPARATIVE DATA ON LATIN TESTS

Table 33 (on page 172) shows the extent to which various Latin tests are measures of the same skills.

Brueckner¹ utilizes four Latin tests (Henmon, Pressey, Tyler-Pressey, and Godsey) in an investigation of Latin skills. Over 5000 high school students took each test, and the study reports quartile scores for students in each semester of Latin from one to eight. Correlations among the tests are

¹ Brueckner, L. J., "The Status of Certain Basic Latin Skills." *Journal of Educational Research*, Vol. XI (May, 1921), pages 390-402.

TABLE 33
LATIN TEST CORRELATIONS

NAME AND FORM OF TEST	<i>r</i>	<i>N</i>
Henmon Vocabulary (1) <i>vs.</i> Henmon Sentence (1)60	47
Henmon Vocabulary (2) <i>vs.</i> Henmon Sentence (2)49	47
Henmon Vocabulary (3) <i>vs.</i> Henmon Sentence (3)57	44
Henmon Vocabulary (4) <i>vs.</i> Henmon Sentence (4)59	43
Stevenson-Coxe Derivatives (1) <i>vs.</i> (2) (Reliability)66	67
Stevenson Vocabulary (1) <i>vs.</i> (2) (Reliability)86	67
Pressey Syntax (Reliability, whole test)78	67
Tyler-Pressey Verb Forms (Reliability, whole test)56	67
Stevenson Vocabulary (1) <i>vs.</i> White Latin (A)47	67
Stevenson-Coxe Derivatives (1) <i>vs.</i> White Latin (A)23	67
Stevenson-Coxe Derivatives (1) <i>vs.</i> Pressey Syntax38	67
Stevenson Vocabulary (1) <i>vs.</i> Tyler-Pressey Verbs47	67
Stevenson-Coxe Derivatives (1) <i>vs.</i> Tyler-Pressey Verbs46	67
Tyler-Pressey Verbs <i>vs.</i> White Latin (A)51	67
Tyler-Pressey Verbs <i>vs.</i> Pressey Syntax43	67
Stevenson Vocabulary (1) <i>vs.</i> Pressey Syntax50	67
Stevenson Vocabulary (1) <i>vs.</i> Stevenson-Coxe Derivatives37	67
Pressey Syntax <i>vs.</i> White Latin (A plus B)38	67

also given. Brueckner's data lead to the following tentative conclusions:

- (1) The Henmon Latin Tests are inadequate for measuring vocabulary and translation beyond four semesters.
- (2) The Pressey Latin Syntax Test and the Tyler-Pressey Test in Latin Verb Forms show little average increase in student ability after the first semester.
- (3) The Godsey Diagnostic Test in Latin Composition shows fairly steady increase from semester to semester in translation and ability to apply rules. It also correlates highest with a composite of the tests used.
- (4) There are large pupil and grade variations in Latin ability, and considerable vagueness as to what constitutes a semester of Latin.

Remedial procedures in Latin. The tests available in Latin represent only a beginning in this field of measurement. The teacher can best meet his needs in Latin, as in French and Spanish, by considering the standard tests supplementary to an extensive series of objective examinations. Helen M. Eddy illustrated a number of useful types of such examinations in Latin (in "A Course of Study in Latin for High Schools," *University of Iowa Extension Bulletin*, No. 112, November 15, 1924). The teacher will find the American Classical League's 1924 Study (*The Classical Investigation*) filled with specific aids in the teaching of Latin, and in securing better coördination between valid aims and measurable results.

Test Materials

FRENCH AND SPANISH

Henmon French Tests. By V. A. C. HENMON. Tests 1, 2, 3, and 4, each test 50 cents per package of 25, including Directions for Administering and Scoring and Record Sheet. Specimen set, 10 cents. World Book Company, Yonkers-on-Hudson, New York.

Handschin Modern Language Tests. By C. H. HANDSCHIN. Silent Reading Test A, French; B, French; A, Spanish; and Comprehension and Grammar Test A, French. Each test \$1.00 per package of 50 sheets, including 4 Record Sheets (with directions). Specimen set, 20 cents. World Book Company, Yonkers-on-Hudson, New York.

Iowa Placement Examinations. French Training, FT-1, Revised (2 forms); Spanish Training, ST-1 (2 forms); and Foreign Language Aptitude, FA-1, Revised (2 forms). Each examination \$3.50 per 100, including Directions and Key. Bureau of Educational Research and Service, University of Iowa, Iowa City, Iowa.

Columbia Research Bureau French Test. By A. A. MÉRAS, SUZANNE ROTH, and BEN D. WOOD. Forms A and B. Each form \$1.30 per package of 25, including Manual of Directions, Key, and Class Record. Specimen set, 20 cents. World Book Company, Yonkers-on-Hudson, New York.

Wilkins Prognosis Test in Modern Languages. By L. A. WILKINS. Each test (8-page booklet containing Tests I-VI and auxiliary 4-weeks tests) \$1.20 per package of 25, including Manual. Specimen set, 10 cents. World Book Company, Yonkers-on-Hudson, New York.

Minimum French Vocabulary Test Book. By C. F. WARD. Lists 2000 commonest words, with French-to-English translations, phonetic transcription, and devices and rules for learning and teaching. 60 cents per copy. The Macmillan Company, New York.

LATIN

Henmon Latin Tests. By V. A. C. HENMON. Tests 1, 2, 3, 4, and X, each test 50 cents per package of 25, including Directions for Administering and Scoring and Record Sheet. Specimen set, 10 cents. World Book Company, Yonkers-on-Hudson, New York.

Ullman-Kirby Latin Comprehension Test. By B. L. ULLMAN and T. J. KIRBY. Forms 1 and 2, each \$1.75 per 100, with Directions. Bureau of Educational Research and Service, University of Iowa, Iowa City, Iowa.

White Latin Test. By D. S. WHITE. Forms A and B, each \$1.20 per package of 25, with Manual of Directions, Key, and Class Record. Specimen set, 20 cents. World Book Company, Yonkers-on-Hudson, New York.

Sterenson-Coxe Latin Derivative Test. By P. R. STEVENSON and W. W. COXE. Forms I, II, and III, each 50 cents per package of 25, with Directions and Key. Specimen set, 15 cents. Public School Publishing Company, Bloomington, Illinois.

Sterenson Latin Vocabulary Test. By P. R. STEVENSON. Forms I, II, and III, each 50 cents per package of 25, with Directions and Key. Specimen set, 15 cents. Public School Publishing Company, Bloomington, Illinois.

Pressey Test in Latin Syntax. By L. W. PRESSEY. 50 cents per package of 25, with Directions and Key. Specimen set, 10 cents. Public School Publishing Company, Bloomington, Illinois.

Tyler-Pressey Test in Latin Verb Forms. By CAROLINE TYLER and S. L. PRESSEY. 50 cents per package of 25, with Directions and Key. Specimen set, 10 cents. Public School Publishing Company, Bloomington, Illinois.

Godsey Latin Composition Test. By EDITH R. GODSEY. Forms A and B, each \$1.00 per package of 25, with Manual of Directions, Key, and Class Record. Specimen set, 15 cents. World Book Company, Yonkers-on-Hudson, New York.

Orleans-Solomon Latin Prognosis Test. By JACOB S. ORLEANS and MICHAEL SOLOMON. Form A, \$1.30 per package of 25, with Manual of Directions, Key, and Class Record. Specimen set, 15 cents. World Book Company, Yonkers-on-Hudson, New York.

References

ALLEN, W. S. "A Study in Latin Prognosis." *Teachers College Contributions to Education*, No. 135 (1923); 41 pages. Columbia University, New York.

BOND, O. F. "Causes of Failure in Elementary French and Spanish Courses at the College Level." *School Review*, Vol. XXXII (1924), pages 276-287.

BRIGGS, T. H. "Lohr-Latshaw Latin Test." *The Classical Journal*, Vol. XVIII (May, 1923), pages 451-465.

— "Prognosis Tests of Ability to Learn Foreign Languages." *Journal of Educational Research*, Vol. VI (December, 1922), pages 386-392.

BRUECKNER, L. J. "The Status of Certain Basic Latin Skills." *Journal of Educational Research*, Vol. IX (May, 1924), pages 390-402.

BYRNE, L. "Latin Tests in Iowa High Schools." *University of Iowa Bulletin*, No. 92 (1923).

CHURCHMAN, P. H. "Courses for Beginners." *The Modern Language Journal*, Vol. IX, No. 4 (January, 1925), pages 207-225.

Classical Investigation, The (Andrew F. West, Chairman). Princeton University Press, Princeton, New Jersey; 1924. 305 pages.

CLEM, ORLIE M. "Latin Prognosis: A Study of the Detailed Factors of Individual Pupils." *Journal of Educational Psychology*, Vol. XVI (March, 1925), pages 160-169.

COLEMAN, ALGERNON. "The First Year of the Modern Foreign Language Study." *The Modern Language Journal*, Vol. X (April, 1926), pages 389-399.

— "What the Modern Language Teacher Must Do for the Modern Foreign Language Study." *The Modern Language Journal*, Vol. X (November, 1925), pages 65-74.

DEIHL, J. D. "The Basis of Educational Tests in Modern Foreign Languages." *The Modern Language Journal*, Vol. VII (1923), pages 269-273.

EDDY, HELEN M. "A Course of Study in French for High Schools." *University of Iowa Extension Bulletin*, No. 105 (May 1, 1924); 64 pages.

— "A Course of Study in Latin for High Schools." *University of Iowa Extension Bulletin*, No. 112 (November 15, 1924); 107 pages.

FIFE, ROBERT H. "Report on the Modern Foreign Language Study in the United States." *Educational Record*, Vol. VI (July, 1925), pages 203-211.

HANDSCHIN, C. H. "Tests and Measurements in Modern Language Work." *The Modern Language Journal*, Vol. IV (1920), pages 217-225.

HENMON, V. A. C. "The Measurement of Ability in Latin." *Journal of Educational Psychology*, Vol. XI (March, 1920), pages 121-136.

— "A French Word Book Based on 400,000 Running Words." *Bureau of Educational Research Bulletin*, No. 3 (1924); 84 pages. University of Wisconsin, Madison, Wisconsin.

HENMON, V. A. C. "Standardized Vocabulary and Sentence Tests in French." *Journal of Educational Research*, Vol. III (1921), pages 81-105.

JORDAN, J. N. "Prognosis in Foreign Language in Secondary Schools." *School Review*, Vol. XXXIII (September, 1925), pages 541-546.

MÉRAS, A. A., ROTH, S., and WOOD, BEN D. "A Placement Test in French." *Contributions to Education*, Vol. I, Chapter XXV. World Book Company, Yonkers-on-Hudson, New York; 1924.

ORTEGA, JOAQUÍN. "Suggestions for Teaching Spanish Conversation and Composition." *The Modern Language Journal*, Vol. VIII (December, 1923), pages 145-158.

RUSSELL, G. OSCAR. "A Silent Reading Test." *The Modern Language Journal*, Vol. IX (May, 1925), pages 459-468.

STARCH, DANIEL. "A Test in Latin." *Journal of Educational Psychology*, Vol. X (December, 1919), pages 489-500.

STODDARD, GEORGE D. "Iowa Placement Examinations." *University of Iowa Studies in Education*, Vol. III, No. 2 (August 15, 1925); 103 pages.

ULLMAN, B. L., and KIRBY, T. J. "A Latin Comprehension Test." *Journal of Educational Research*, Vol. X (November, 1924), pages 308-317.

WARD, C. F. *Minimum French Vocabulary Test Book* (2000 commonest words). The Macmillan Company, New York; 1926. 101 pages.

WEST, ANDREW F. (Chairman). *The Classical Investigation*. Princeton University Press, Princeton, New Jersey; 1924. 305 pages.

WILKINS, L. A. "Results in a Prognosis Test Given to Pupils Beginning French and Spanish." *Bulletin of High Points*, Vol. I, No. 8 (October, 1919), pages 26-30. Board of Education, New York.

— "Problems in the Modern Language Field and Attempted Solutions." *Contributions to Education*, Vol. I, Chapter XXIV. World Book Company, Yonkers-on-Hudson, New York; 1924.

WOODY, C. *The Ullman-Kirby and Godsey Latin Tests and the Carr English Vocabulary Test*. Bulletin No. 56 (May 21, 1923). Bureau of Educational Reference and Research, University of Michigan, Ann Arbor, Michigan.

— *Report of Latin Investigation in Various High Schools of Michigan*. Bulletin No. 64 (March 31, 1924). Bureau of Educational Reference and Research, University of Michigan, Ann Arbor, Michigan.

YOUNG, C. E. "French and Spanish in the High School." *University of Iowa Extension Bulletin*, No. 93 (September 1, 1923); 29 pages.

CHAPTER NINE

SOCIAL STUDIES

Introduction. High school teachers of the social studies will recall the recommendations in the 1916 Report of the Committee on Social Studies.¹ The courses proposed for high school, under either a 6-3-3 or 8-4 plan, were the following:

Year IX.	(1) Community Civics ($\frac{1}{2}$ year)	
	Civics — economic and vocational ($\frac{1}{2}$ year)	
	History	
or	(2) Civics — economic and vocational	{ 1 year, in sequence or parallel
Years X-XII.	Economic History	
	European History to end of 17th Century (1 year)	
	European History (including English History) since 17th Century (1 or $\frac{1}{2}$ year)	
	American History since 17th Century (1 or $\frac{1}{2}$ year)	
	Problems of American Democracy (1 or $\frac{1}{2}$ year)	

Allowance was made for community and individual needs, and flexibility was urged in building up the program of study. More recently courses have been recommended as follows:²

- Year IX. Community and National Activities
- Year X. Modern European History since 1650
- Year XI. American History during the National Period
- Year XII. Problems of Democracy

¹ Dunn, A. W. (Secretary of Committee), *The Social Studies in Secondary Education*, Bulletin No. 28 (1916); 63 pages (Department of the Interior, Bureau of Education, Washington, D. C.).

² See Rugg, Earle, *The Twenty-second Yearbook of the National Society for the Study of Education*, Part II, pages 62-63.

The Twenty-second Yearbook, Part II,¹ was devoted to an investigation of the social studies and is noteworthy for its definite turning from the encyclopedic point of view to that of problem-solving, from formal knowledge to active projects drawing at once upon geography, history, and civics. The field covered by social science is extremely complex and ordinarily only loosely organized. There is still little agreement on the relative merits of the special subjects and on the relative weight and sequence of their instructional units. Hence it is again necessary to point out (1) that adequate standard tests have not appeared in the group of social studies and (2) that such tests cannot be developed until instructional units, courses of study, methods, and aims are better understood and definite programs become more universally established. For example, factual tests cover only material which is generally recognized as least valuable; and grade norms for tests are meaningless where sequences vary. Only some of the more recent tests afford brief measures of the socially valuable phases of ability in history.

Barr Diagnostic Tests in American History

Description of the tests. Five tests, printed in an 8-page folder, cover (1) Comprehension, (2) Chronological Judgment, (3) Historical Evidence, (4) Evaluation of Facts, and (5) Causal Relationships. The tests place considerable emphasis on factual material (although such is not their intent), and they have not been adequately standardized. However, their chief defect is their inadequacy for pupil diagnosis in the five fundamental categories of historical ability. Six minutes only is allowed for each test, and the number of pupil-reactions obtained in each case is small. Scoring is unnecessarily complicated by means of a weighting system.

¹ *Op. cit.*; 324 pages.

Interpretation and utilization of results. The norms given for the Barr Diagnostic Tests in American History are not of great value, since they are simply grade medians for an unstandardized test; and the test should not be looked upon as a measure of achievement in history. Nevertheless, the test is of value in pointing out what is important in history-ability and illustrating practicable methods of objective testing in a difficult field. It is recommended that a history teacher employing these tests give both series (2 A and 2 B) and in addition construct similar examinations more specifically related to modern American history. The World War, it may be noted here, is scarcely mentioned in history tests.

Validity and reliability. The Barr Diagnostic Tests in American History must be considered as only a brief sampling of a restricted period. The units selected for diagnostic measurement appear important, but the data of Table 34 indicate that the separate tests are too brief.

TABLE 34

DATA ON THE RELIABILITY OF THE BARR DIAGNOSTIC TESTS IN
AMERICAN HISTORY

TEST	<i>r</i>	<i>N</i>	S.D. 2A	S.D. 2B	P.E. _{score}	P.E. _{.1}	P.E. _{score} S.D.	P.E. _{.1} S.D.	NATURE OF GROUP
1	.63	50	3.3	3.2	1.3	1.1	.40	.33	Gr. IX and XII
2	.53	50	2.3	2.3	1.1	.8	.47	.34	Gr. IX and XII
3	.30	50	1.3	2.0	.9	.5	.53	.31	Gr. IX and XII
4	.24	50	4.2	3.3	2.2	1.1	.58	.29	Gr. IX and XII
5	.62	50	3.0	2.6	1.2	.9	.43	.34	Gr. IX and XII
Total	.77	50	10.7	8.7	3.1	2.7	.32	.28	Gr. IX and XII

Pressey-Richards American History Test

The Pressey-Richards Test for Understanding American History covers:

- (1) Character Judgment (25 items, 5 minutes)
- (2) Historical Vocabulary (25 items, 6 minutes)
- (3) Sequence of Events (25 items, 6 minutes)
- (4) Cause and Effect Relationships (25 items, 8 minutes)

It is designed to provide a brief measure of the abilities listed, but is primarily factual. "Character Judgment" consists in finding the best single adjective which applies to well-known personages: the correct answer is usually an uncritical generalization. The "Historical Vocabulary" is useful but restricted to pre-World War terms. It could be readily extended by the teacher of history. The test on "Sequence of Events" is filled with implied dates considered socially unimportant in recent investigations (e.g., *The Twenty-first Yearbook*), and the test on "Cause and Effect Relationships" offers little opportunity for thoughtful reaction on the part of the student. Table 35 gives data on the reliability of the test.

Gregory Tests in American History, Test III

Description of the test. Test III of the Gregory Tests in American History is designed to measure history for Grades VIII to XII inclusive. The following names of the parts indicate the contents of the test:

- Part 1. Miscellaneous Facts and Dates
- Part 2. The Period of Discovery, Exploration, and Colonization
- Part 3. The Period of Revolution, from 1760 to 1789
- Part 4. The Period of National Growth, from 1789 to 1830

- Part 5. The Period of Sectional Disputes and Civil War, 1830 to 1865
- Part 6. The Period of Reconstruction and National Development, from 1865 to 1900
- Part 7. The Period from 1900 to 1922

There are 10 questions under each part except the first, which has 40 questions.

Interpretation and utilization of results. Although 40 of the 100 items in this test are designed to measure the more fundamental phases of history, they are really primarily factual. The brevity of each of the parts necessitated but scanty treatment of historically important periods. Thus in Part 7, which covers the most recent period in American History, but 2 questions out of the 10 are connected with the World War.

For data on the reliability and validity of these tests, see Tables 35 and 36.

The Van Wagenen Reading Scale in History (Forms A and B) is useful for a measure of comprehension, but it involves a noticeable dependence upon memorized historical facts. One scale (Information S-3) in the Van Wagenen American History series is designed for high school use. (See Tables 35 and 36.)

The Kepner Background Tests in Social Sciences (Forms A and B) diagnose pupil weaknesses in factual material, as follows:

- Exercise I. Association of Men and Events
- Exercise II. Literary Background
- Exercise III. Geographic Concepts
- Exercise IV. Historical Vocabulary
- Exercise V. Social and Economic Vocabulary
- Exercises VI and VII. Dates and Chronology

The utility of these tests is lessened by reason of the diversity of pupil-preparation in the informations tested and the lack of existing standards in the teaching of history; who can say what background is essential? These tests have, however, been prepared with considerable care and show a fair degree of reliability. (See Table 35.) There is some question whether or not the Kepner Tests differ greatly from "pure" achievement tests in the functions measured. (See Table 36.)

Brown-Woody Civics Test

Description of the test. The Brown-Woody Civics Test consists of three parts designed to measure, respectively, civic vocabulary, civic information, and civic thinking. The working time is 35 minutes.

Interpretation and utilization of results. Tentative percentile norms for pupils in junior and senior high school are given in the Manual of Directions. The test should prove serviceable in measuring pupil and class progress, and in class sectioning.

Validity and reliability. Each item in the test covers a point brought out in at least five of nine common textbooks in civics. The Brown-Woody Civics Test has a reliability of .92.

Remedial procedures in social studies. The tests mentioned have been largely historical in content, and of the factual type. They are typical of those available for high school teachers in the social studies. Results obtained with them can form a basis for rough classification — for example, in the apportionment of tasks in a project or socialized recitation. Such tests cannot measure the larger purposes of the social studies nor can objective examinations covering similar elements. Such aims as changed attitudes, enlarged understanding of, and interest in, social needs and responsibilities may be amenable to standard measurement, but first must come improved teaching and testing technique.

TABLE 35¹
RELIABILITIES OF CERTAIN HISTORY TESTS

	GREGORY	BARR	PRESSEY-RICHARDS	VAN WAGENEN S-3	KEPNER	VAN WAGENEN SCALE R
<i>r</i>	.79	.71	.89 ²	.76 ²	.79	.57
<i>N</i>	290	279	296	225	215	217
Mean (1) ³	38.7	47.5	56.2	75.5	48.3	82.1
Mean (2) ³	41.8	48.8	—	—	50.5	82.3
S.D. (1) ³	16 ⁴	12	14	8	8	7
S.D. (2) ³	16	12	—	—	8	7

TABLE 36

INTERCORRELATIONS OF A NUMBER OF HISTORY TESTS BASED UPON 240 HIGH SCHOOL JUNIORS AND SENIORS, MASON CITY, IOWA (*Data from same source as Table 35*)

TEST	1	2	3	4	5	6	7	8	9	10	AVER-AGE
1. Gregory A . .	—	.79	.59	.69	.72	.68	.76	.71	.49	.51	.66
2. Gregory B . .	.79	—	.52	.61	.72	.70	.72	.69	.49	.45	.63
3. Barr 2 A . .	.59	.52	—	.71	.61	.52	.60	.53	.35	.44	.54
4. Barr 2 B . .	.69	.61	.71	—	.69	.57	.65	.56	.53	.52	.61
5. Pressey-Richards . .	.72	.72	.61	.69	—	.67	.82	.71	.56	.55	.67
6. Van Wagenen S-3 . .	.68	.70	.52	.57	.67	—	.72	.66	.45	.44	.60
7. Kepner A . .	.76	.72	.60	.65	.82	.72	—	.79	.54	.47	.67
8. Kepner B . .	.71	.69	.53	.56	.71	.66	.79	—	.38	.40	.60
9. Van Wagenen Hist. Read. A .	.49	.49	.35	.53	.56	.45	.54	.38	—	.57	.48
10. Van Wagenen Hist. Read. B .	.51	.45	.44	.52	.55	.44	.47	.40	.57	—	.48

¹ Ruch, G. M., et al. *Objective Examination Methods in the Social Studies*, Chapter VI (Scott, Foresman & Co., 1926). A report of an investigation under a grant from the New York Commonwealth Fund.

² Computed by method of "odds" vs. "evens," stepped up by the Spearman-Brown formula. All other reliabilities are form vs. form.

³ Mean (1) refers to the first form, and Mean (2) to a second form. The same designation is used for the standard deviations.

⁴ To the nearest integral values.

Test Materials

Barr Diagnostic Tests in American History. By A. S. BARR. Series A and B, each series \$4.00 per 100 (or 5 cents each in small quantities). Specimen set, 20 cents. Public School Publishing Company, Bloomington, Illinois.

Pressey-Richards American History Test (Understanding of American History). By L. W. PRESSEY and R. C. RICHARDS. \$2.00 per 100. Specimen set, 10 cents. Public School Publishing Company, Bloomington, Illinois.

Gregory Tests in American History. By C. A. GREGORY. Test III, Forms A and B, each form \$1.00 per package of 25. Specimen set, 10 cents. Bureau of Administrative Research, University of Cincinnati, Cincinnati, Ohio.

Van Wagenen Reading Scales — History. By M. J. VAN WAGENEN. Scales A and B, each \$3.00 per 100. Specimen set, 20 cents. Public School Publishing Company, Bloomington, Illinois.

Van Wagenen History Scales, Information Scale S-3. By M. J. VAN WAGENEN. \$2.00 per 100, including Directions, Key, and Record Sheet. Bureau of Publications, Teachers College, Columbia University, New York.

Kepner Background Test in Social Sciences. By TYLER KEPNER. Forms A and B, each \$1.25 per package of 25, with Directions and Key. Specimen set, 25 cents. Harvard University Press, Cambridge, Massachusetts.

Brown-Woody Civics Test. By A. W. BROWN and CLIFFORD WOODY. Form A, \$1.30 per package of 25, with Manual of Directions, Key, and Class Record. Specimen set, 15 cents. World Book Company, Yonkers-on-Hudson, New York.

References

BRINCKLEY, S. G. "Values of New-Type Examinations in the High School, with Special Reference to History." *Teachers College Contributions to Education*, No. 161 (1924); 121 pages. Columbia University, New York.

DAWSON, E. "Organizing the Social Studies." *Contributions to Education*, Vol. I, Chapter XVIII. World Book Company, Yonkers-on-Hudson, New York; 1924.

DUNN, A. W. *The Social Studies in Secondary Education*. Report of the Committee on Social Studies of the Commission on the Reorganization of Secondary Education of the National Education Association. Bulletin No. 28 (1916). Department of the Interior, Bureau of Education, Washington, D. C.

HENMON, V. A. C. "Some Limitations of Educational Tests." *Journal of Educational Research*, Vol. VII (March, 1923), pages 185-198.

KEPNER, P. T. "A Survey of the Test Movement in History." *Journal of Educational Research*, Vol. VII (April, 1923), pages 309-325.

ODELL, C. W. "The Barr Diagnostic Tests in American History." *School and Society*, Vol. XVI (October 28, 1922), pages 501-503.

RUCH, G. M. *Objectite Examination Methods in the Social Studies*. Scott, Foresman & Co., Chicago; 1926.

RUGG, H. O. "The Social Studies in the Elementary and Secondary School." *The Twenty-second Yearbook of the National Society for the Study of Education* (1923); 324 pages.

VAN WAGENEN, M. J. "Revised Van Wagenen History Scales." *Teachers College Record*, Vol. XXVII, No. 2 (October, 1925), page 142.

CHAPTER TEN

VOCATIONAL SUBJECTS

I. MECHANICAL AND TRADE TESTS

Introduction. Measurement in the various subjects usually designated, somewhat loosely, "vocational subjects" has progressed unevenly. Industrial concerns have been glad to avail themselves of the material developed during the war for the purpose of classifying tradesmen, and they have used a few employment and clerical tests to advantage. But standardized tests for all the subjects in the commercial departments of high schools, and for such highly specialized skills as music and art, have been slow to appear. The pioneer work in music consisted of the tests for musical talent or aptitude developed by C. E. Seashore, and they remain the outstanding measuring instrument in this field. Tests recently constructed by Kwalwasser and Ruch mark the beginning of standardized measurement of musical accomplishment — i.e., of the pupil's knowledge of music.

Stenquist Mechanical Aptitude Tests

Description of the tests. Test I of the Stenquist Mechanical Aptitude Tests consists of 95 pictorial items covering common mechanical objects. The student is asked to indicate which object "belongs with, is used with, or is a part of" another object. In Test II, 78 items are presented which call for mechanical perception and abstract reasoning.

What the tests measure. The Stenquist tests may be said to measure general mechanical ability or aptitude. Stenquist has shown that this aptitude is not primarily a function of the student's definite mechanical training, but that it is distributed among children much as general in-

telligence is. That it is largely independent of general intelligence is shown by the low correlations (.2 to .4) obtained between these tests and standard intelligence tests. The median correlation obtained between the two tests and instructors' ratings in 15 classes in science and shopwork was .67. The same relation was found between the two tests and the Stenquist Assembling Tests, which call for actual mechanical manipulation on the part of the student.

Administration and scoring. Giving the Stenquist tests presents no difficulties. Test I requires 45 minutes' working time, and Test II, 50 minutes. It is recommended that both tests be given in every case. Scoring is objective and can be accomplished very readily.

Interpretation and utilization of results. Scores in the Stenquist Mechanical Aptitude Tests have special significance in conjunction with results from intelligence tests. As will be pointed out in Chapter XII, the latter, by themselves, have only limited bearing upon the student's vocational capacities and needs. Students standing high in intelligence and in mechanical aptitude may be expected to become the leading engineers; but there is considerable opportunity for success on the part of students possessing only unusual mechanical ability. From this class particularly industry can recruit its ranks of skilled technicians; and the individuals in turn can, through early training along the lines of special capacities, best reach their highest utilitarian and social development.

Validity and reliability. The validity of the Stenquist tests rests in part upon the correlations with teachers' judgments, already mentioned, and in part upon the low correlation with intelligence. It should be noted that the tests are general in nature, and do not give a measure of skill in any particular trade. Correlations obtained between the Stenquist tests and standard intelligence tests are uniformly low,

thus indicating that different functions are being measured. Stenquist has reported to the authors a correlation of .68 between Tests I and II on 230 boys in Grades V to VIII in New York City. Further data on the reliability of these tests are given in Table 37. The test is clearly more effective for boys than for girls.

TABLE 37

DATA ON THE RELIABILITY OF THE STENQUIST MECHANICAL APTITUDE TESTS

<i>r</i>	<i>N</i>	S.D.	P.E. _{age}	P.E. ₁	P.E. _{score} S.D.	P.E. ₁ S.D.	NATURE OF GROUP
.58	66	8.9	3.9	3.0	.43	.33	Grades VII-XII (Girls)
.84	79	12.4	3.4	3.1	.27	.25	Grades VII-XII (Boys)
.97	45	18.1	2.1	2.1	.11	.11	Grades VII-XII (Boys and Girls)

Norms. Percentile age-norms in one-year increments from age 11 years, 6 months to 15 years, 6 months are given in the Manual of Directions. Equivalent *T*-scores are given also, all values being tabulated separately for Test I and Test II. As a matter of fact, there is little increase in median ability from year to year. The important thing to know is the student's standing relative to other individuals of the same age. Bell¹ has shown that the medians for men run from 10 to 15 points higher than those for women.

Thurstone Vocational Guidance Tests

Description of the tests. The six tests of the Thurstone Vocational Guidance series comprise the following: Arithmetic, Algebra, Geometry, Physics, Technical Information, and

¹ Bell, J. Carleton, "Mechanical Aptitude and Intelligence." *Contributions to Education*, Vol. I, Chapter XXVII (World Book Company, 1924).

the Thurstone Psychological Examination. The Technical Information Test covers "such technical information as a boy would acquire from reading popular technical journals, constructing mechanical toys, inquiring about automobile engines," etc. It is similar to the other tests in the series, which have been described briefly in the chapters devoted to particular subjects.

Interpretation and utilization of results. Data on the reliability of the various tests are not available, but the diagnostic power for freshman engineering scholarship (which is the chief purpose of the series) is not great. Median coefficients of correlation between first-year engineering pooled grades and the various Thurstone Vocational Guidance Tests range from .23 for Technical Information to .42 for Algebra. The battery of six tests gave a median correlation with pooled first-year grades of .45. However, the tests are more useful in such prediction than are high school grades in specific subjects. Effective prediction of the success of a particular student in engineering will demand a type of measurement which includes mental, educational, character, and environmental factors. Also, the criterion of success — i.e., college marks — must first be greatly improved in reliability and validity.

Trade Tests

Description of the tests. Trade tests afford a measure of skill in a particular trade through utilization of test material drawn from that trade. The technique developed during the World War by the Committee on Classification of Personnel has formed the basis for more recent standardizations of trade test procedures.

Chapman has shown clearly how trade tests differ from intelligence tests and "skill prediction tests." He states:¹

¹ Chapman, J. C., *Trade Tests* (Henry Holt & Co., 1921; 374 pages).

The trade test makes no pretense at measuring intelligence directly; it makes no attempt to measure the native endowment of the subject, with a view to predicting the degree of success to be expected as a result of training in a specific subject; the trade test furnishes a rating, in objective quantitative terms, of the degree of trade ability already possessed as a result of practice in the trade. Nevertheless, the trade rating can, under certain conditions, be used as a help in predicting the future capability of a tradesman.

Toops¹ states that standardized trade tests are destined to play an important rôle in vocational counseling in the high school, and that scientific vocational placement depends upon "(a) knowledge of the pupil's abilities, and (b) the requirements of the job, in the way of human abilities and acquirements." It is evident that only tests developed specifically for this purpose can give a valid measure of an individual's vocational fitness. One's performance in a general intelligence test establishes certain broad divisions of native talent, but the problem for vocational schools is that of differentiating among pupils in approximately the same intelligence categories.

Uses of trade tests in vocational guidance. It is beyond the scope of this book to evaluate the construction and utilization of such tests, inasmuch as this information would not primarily concern teachers in most high schools. But trade tests merit special attention on the part of teachers in vocational and technical high schools and of vocational counselors, who will find the studies by Chapman and Toops of considerable interest and value, as they contain many useful tests, with complete directions for scoring and for interpretation of results. For further knowledge the following publications are recommended:

¹ Toops, H. A., "Trade Tests in Education." *Teachers College Contributions to Education*, No. 115 (1921), pages 2-4 (Columbia University, New York).

GRIFFITS, C. H., *Fundamentals of Vocational Psychology* (Macmillan, 1924).
LINK, H. C., *Employment Psychology* (Macmillan, 1919).
RUGGLES, A. M., "A Diagnostic Test of Aptitude for Clerical Office Work" (Teachers College, Columbia University, 1924).
TOOPS, H. A., *Tests for Vocational Guidance of Children Thirteen to Sixteen* (Teachers College, Columbia University, 1923).

II. COMMERCIAL TESTS

Introduction. Commercial tests of a type suitable for high school group measurement are still in the experimental stage. The commercial field involves a great number of aptitudes and skills, many of which can be measured indirectly by standard mental and educational tests, such as tests of intelligence, English grammar, spelling, punctuation, and arithmetic. The Thurstone Clerical Examination, the Thurstone Proficiency Test for Typists, and the Cody Commercial Tests are examples of tests which possess utility in business and industry but are not adequate for the needs of the commercial department of the modern high school. The Cody monograph (*Commercial Tests and How to Use Them*, World Book Company) includes a complete series of measurements of clerical ability which may serve as a starting point for teachers desirous of building up a battery of informal, objective commercial tests.

Blackstone Stenographic Proficiency Tests, Typewriting

Description of the tests. The Blackstone Typewriting Test is available in five equivalent forms. The points of approximate equivalence of the five business letters comprising the different forms are the following:

- (1) Number of words
- (2) Number of letters
- (3) Number of letters struck with each hand
- (4) Number of carriage returns and shift-key strokes
- (5) Position of long or difficult words in the text

Administration and scoring. Specific directions are given for controlling conditions, so that pupils taking the test will have their performance measured only in terms of speed and accuracy. The stroke method of counting is employed. The working time of the test is 3 minutes, but 5 minutes' practice on other material is a required fore-exercise. Pupils may exchange papers and score the tests themselves in accordance with the explicit directions given.

Interpretation and utilization of results. The Blackstone Typewriting Test is useful in finding the general class level after various periods of instruction, and in motivating the individual pupil. The latter service is best rendered by recording graphically the pupil's progress in speed and accuracy. Norms for various periods of instruction in intervals of five months are given in the Manual of Directions.

III. MUSIC TESTS

Seashore Measures of Musical Talent

Description of the tests. The Seashore Measures of Musical Talent consist of six Columbia phonograph records, as follows:

- A 7536 Sense of Pitch, No. 1A and No. 1B
- A 7537 Sense of Intensity, No. 2A and No. 2B
- A 7538 Sense of Time, No. 3A and No. 3B
- A 7539 Sense of Consonance, No. 4A and No. 4B
- A 7540 Sense of Memory, No. 5A and No. 5B
- 53005-D Sense of Rhythm, No. 6A and No. 6B

All records are of the standard 12-inch type, and can be played on any first-class phonograph. No other material need be bought, but a score card similar to the model given in the Manual of Directions should be prepared for each pupil.

What the tests measure. The Seashore tests measure six basic capacities which underlie general music talent. Pitch, intensity, time, and rhythm are measured in terms of least perceptible difference; consonance, in ability to judge degree of consonance or dissonance; and tonal memory in terms of memory span for a sequence of unrelated tones.

Administration and scoring. Seashore recommends that the tests be given first in the fifth grade, in order to enable early plans for the musical education of some children, and again in the eighth grade, since vocational interests are paramount during that period. However, they can be given in any high school grade and to adults.

Any teacher can administer the tests (or they can be given in the home) by following rigorously the directions accompanying the records. Each test can be given at one time to all the pupils in a single classroom.

The obtained score (number of correct judgments) is reduced to per cent right by dividing by the total number of trials, and this in turn is transformed into a percentile rank in accordance with tables in the Manual of Directions. The highest rank (i.e., corresponding to the highest score actually obtained in a large number of trials) is given a rank of 100, the mean, 50, and the lowest, 1. Norms (percentile ranks) are given for Grades V and VIII, and for adults. High school grade-norms are found by interpolation.

Interpretation and utilization of results. Seashore¹ gives the following rules for musical guidance on the basis of capacity for pitch discrimination:

Other things being equal, and due allowance having been made for this record (pitch) in relation to all other records of musical capacity, the advice should be as follows:

¹ Seashore, C. E., *The Psychology of Musical Talent* (Silver, Burdett & Co., 1919), pages 67-68.

Best 10 per cent: encourage freely;
Next 20 per cent: encourage;
Next 40 per cent: question;
Next 10 per cent: discourage.

He recommends also that counsel be given, where possible, in the light of a complete survey of the pupil's musical capacity. Scores in the various tests are not to be averaged, since that procedure would obscure significant facts.

Norms. Norms of the type referred to above are given for all tests in the Manual of Directions, and more complete descriptions of pupil-performance will be found in C. E. Seashore, *The Psychology of Musical Talent*. For effective use the attainment of a pupil is plotted graphically on a chart which includes the norm for each trait. The amounts of departure above or below the norm for each trait, and the aggregate of musical talent indicated by the whole profile, must be taken into account in interpreting the results obtained.

Validity and reliability. The Seashore Musical Talent Tests cover only a limited range of the whole field of musical ability; they simply supply the parent, teacher, or music supervisor with certain objective measurements which increase the likelihood of proper musical guidance. Data on reliability are given in Table 38. Table 39 indicates the degree of relationship between different factors in musical aptitude as measured by these tests.

Kwalwasser-Ruch Test of Musical Accomplishment

Description of the test. Ten parts, printed in an 8-page folder, comprise the following material:

- (1) Knowledge of Musical Symbols and Terms
- (2) Recognition of Syllable Names
- (3) Detection of Pitch Errors in a Familiar Melody

TABLE 38

DATA ON THE RELIABILITY OF THE SEASHORE MEASURES OF MUSICAL TALENT

TEST	<i>r</i>	N	S.D.	P.E-score	P.E. _{±.1}	P.E. _{score} S.D.	P.E. _{±.1} S.D.
Pitch70	100	11.95	4.4	3.7	.37	.31
Intensity . .	.66	100	8.12	3.2	2.6	.40	.32
Time53	100	7.86	3.7	2.7	.46	.34
Consonance . .	.35	100	7.71	4.2	2.5	.54	.32
Memory . .	.66	100	15.30	6.0	4.9	.40	.32
Rhythm50	58	7.22	3.4	2.4	.48	.34

TABLE 39

INTERCORRELATIONS AMONG SEASHORE MEASURES OF MUSICAL TALENT¹
(Based on 210 College Students)

	INTENSITY	TIME	CONSONANCE	TONAL MEMORY
Pitch32	.30	.78	.52
Intensity . . .	—	.23	.24	.20
Time	—	—	.48	.28
Consonance . .	—	—	—	.75

- (4) Detection of Time Errors in a Familiar Melody
- (5) Recognition of Pitch Names
- (6) Knowledge of Time Signatures
- (7) Knowledge of Key Signatures
- (8) Knowledge of Note Values
- (9) Knowledge of Rest Values
- (10) Recognition of Familiar Melodies from Notation

¹ Weaver, A. T., "Experimental Studies in Vocal Expression." *Journal of Applied Psychology*, Vol. VIII (1924), page 171.

Each test is completely objective. The test measures musical accomplishment from Grade IV to Grade XII, inclusive.

Administration and scoring. The Kwalwasser-Ruch test can be easily given by the classroom teacher. Time limits for the sub-tests range from 3 to 8 minutes. Scoring is done rapidly by means of a key. No training in music or in testing is necessary for administration and scoring.

Interpretation and utilization of results. A pupil's score in the Kwalwasser-Ruch test gives a reliable indication of his achievement in public school music. The test can be used for determining school, class, and pupil standing, and, to some extent, diagnostically in showing where the pupil weaknesses lie.

Validity and reliability. The validity of the Kwalwasser-Ruch Test of Musical Accomplishment rests principally upon recommendations unanimously adopted by the Music Supervisors' National Conference, supplemented by studies of courses in several cities prominent for their work in public school music. The reliability of the various sub-tests, as computed from 167 pupils in Grades VI, VIII, X, and XII, ranges from .70 to .95; and the reliability of the whole test, for this group, was found to be .97. The probable error of a score was 6 score points.

Norms. Grade norms are given in Table 40. In addition decile standards are tabulated in the Manual of Directions.

TABLE 40

NORMS FOR THE KWALWASSER-RUCH TEST OF MUSICAL ACCOMPLISHMENT

GRADE	IV	V	VI	VII	VIII	IX	X	XI	XII
Mean . .	72.3	86.1	105.1	120.7	133.7	150.5	163.1	171.4	175.1
Number of Cases	421	519	539	702	665	314	183	114	130

Remedial procedures in music. The music supervisor, no less than the classroom teacher, must take into account individual differences in capacity and accomplishment. The musical accomplishment test described above provides for the first time a standard test of high validity and reliability, in the light of which music teachers and supervisors may, with assurance, evaluate performance in public school music. At the same time valuable information will be gained as to which of the important instructional units need further emphasis.

Test Materials

Stenquist Mechanical Aptitude Tests. By J. L. STENQUIST. Tests I and II, each \$1.50 per package of 25, including Key and Record Sheet. Manual of Directions, 15 cents. Specimen set, 30 cents. World Book Company, Yonkers-on-Hudson, New York.

Measurements of Mechanical Ability. By J. L. STENQUIST. 101 pages. Cloth, \$1.75; paper, \$1.25. Bureau of Publications, Teachers College, Columbia University, New York.

Thurstone Vocational Guidance Tests. By L. L. THURSTONE. Each of the five tests (Arithmetic, Algebra, Geometry, Physics, and Technical Information) \$1.00 per package of 25, with Key and Record Sheet. Manual of Directions, 20 cents. Specimen set, 40 cents. World Book Company, Yonkers-on-Hudson, New York.

Trade Tests. By J. C. CHAPMAN. Cloth-bound, 435 pages. \$4.00. Henry Holt & Co., New York.

Trade Tests in Education. By H. A. TOOPS. 118 pages. Cloth, \$2.00; paper, \$1.50. Bureau of Publications, Teachers College, Columbia University, New York.

Tests for Vocational Guidance of Children Thirteen to Sixteen. By H. A. TOOPS et al. 169 pages. Cloth, \$1.60; paper, \$1.25. Bureau of Publications, Teachers College, Columbia University, New York.

Blackstone Stenographic Proficiency Tests. By E. G. BLACKSTONE. Typewriting, Forms A, B, C, D, and E, each \$1.00 per package of 25, including Manual of Directions, Percentile Graph, and Class Record. Specimen set, 25 cents. World Book Company, Yonkers-on-Hudson, New York.

Commercial Tests and How to Use Them. By SHERWIN CODY. A monograph of 216 pages. \$1.20. World Book Company, Yonkers-on-Hudson, New York.

Measures of Musical Talent. By C. E. SEASHORE. Six phonograph records, each \$1.25. Order according to the numbers and titles given in this chapter. Columbia Graphophone Company, New York.

Kwalwasser-Ruch Test of Musical Accomplishment. By J. KWALWASSER and G. M. RUCH. \$5.00 per 100, \$40.00 per 1000, including Key and Manual of Directions. Extension Division, University of Iowa, Iowa City, Iowa.

References

BELL, J. CARLETON. "Mechanical Aptitude and Intelligence." *Contributions to Education*, Vol. I, Chapter XXVII. World Book Company, Yonkers-on-Hudson, New York; 1924.

Business Education in Secondary Schools. Bulletin No. 55 (1919); 68 pages. Department of the Interior, Bureau of Education, Washington, D. C.

CHAPMAN, J. C. *Trade Tests.* Henry Holt & Co., New York; 1921. 435 pages.

CHURCH, C. *A Survey of Public School Music, Grades IV to XII.* Thesis. University of Iowa, Iowa City; 1926.

CODY, S. *Commercial Tests and How to Use Them.* World Book Company, Yonkers-on-Hudson, New York; 1919. 216 pages.

DYKEMA, PETER. "Tests and Measurements in Music Education." *Proceedings of the National Association of Music Teachers*; 1925.

GRIFFITHS, C. H. *Fundamentals of Vocational Psychology.* The Macmillan Company, New York; 1924. 372 pages.

KLATER, N. J. *The Effect of Training in Rhythm upon Rhythm Discrimination in the Intermediate Grades.* Thesis. University of Iowa, Iowa City; 1924.

KWALWASSER, J. "The Measurement of the Sense of Rhythm." *Musical Observer* (June, 1924).

— "Scientific Testing in Music." *Proceedings of the National Association of Music Teachers*; 1925.

LINK, H. C. *Employment Psychology.* The Macmillan Company, New York; 1919. 440 pages.

RUGGLES, A. M. "A Diagnostic Test of Aptitude for Clerical Office Work." *Teachers College Contributions to Education*, No. 148 (1924); 93 pages. Columbia University, New York.

SCHOEN, MAX. "Common Sense in Music Testing." *Proceedings of the National Association of Music Teachers*; 1925.

— "Tests of Musical Feeling and Musical Understanding." *Journal of Comparative Psychology*, Vol. V (February, 1925), pages 31-52.

SEASHORE, C. E. *The Psychology of Musical Talent.* Silver, Burdett & Co., New York; 1919. 288 pages.

— "A Survey of Musical Talent in the Public Schools." *University of Iowa Studies in Child Welfare*, Vol. I, No. 2 (1920); 36 pages.

SNEDDEN, DAVID. *Vocational Education.* The Macmillan Company, New York; 1920. 587 pages.

STENQUIST, J. L. "A Case for the Low IQ." *Journal of Educational Research*, Vol. IV (November, 1921), pages 241-254.

THURSTONE, L. L. "Intelligence Tests for Engineering Students." *Journal of Engineering Education*, Vol. XIII (1923), pages 263-318.

TOOPS, H. A. "Tests for Vocational Guidance of Children Thirteen to Sixteen." *Teachers College Contributions to Education*, No. 136 (1924); 159 pages. Columbia University, New York.

— "Trade Tests in Education." *Teachers College Contributions to Education*, No. 115 (1921); 118 pages. Columbia University, New York.

VITELES, M. S. "Psychological Tests in Guidance: Their Use and Abuse." *School and Society*, Vol. XXII, No. 560 (September 19, 1925), pages 350-356.

Vocational Guidance in Secondary Education. Bulletin No. 19 (1918); 29 pages. Department of the Interior, Bureau of Education, Washington, D. C.

WEAVER, A. T. "Experimental Studies in Vocal Expression." *Journal of Applied Psychology*, Vol. VIII (1924), pages 23-51, 159-186.

CHAPTER ELEVEN

SURVEY TESTS

Introduction. A few tests which have been widely used by colleges are of such nature that they afford a useful means of measuring and evaluating the work of high schools. For elementary and intermediate grades and the junior high school, the measurement of general school achievement can best be accomplished by means of the Stanford Achievement Test (see Chapter XIII).

There are, of course, some tests not described here which will be of occasional use for survey purposes in high schools; but appropriate data on their reliability and utility are not available.

Iowa Comprehension Test

Description of the test. The Iowa Comprehension Test measures reading comprehension ability of the type most needed for success in college work. Three paragraphs cover abstract material in science, history, and literature. Questions are asked about the text. The student must grasp logical relationships in order to get the correct answer. The factor of guessing is practically eliminated, and scoring is made completely objective by the device of writing down only the number of the bracket which contains the correct answer. The test contains 45 items and requires 36 minutes' working time. A fore-exercise is provided.

Interpretation and utilization of results. The test can best be given during the last month of the senior year of high school. The principal or vocational counselor will find the rank of individual students useful in giving advice in regard to further schooling. Poor performance in any single group test should not weigh heavily in making educational decisions; but extremely high performance in the Iowa

Comprehension Test is usually indicative of ability to do college work.

Validity and reliability. The validity of the Iowa Comprehension Test rests upon its power to predict the character of the work a student will do in college, particularly in the freshman year. Computations made at the University of Iowa show that the Iowa Comprehension Test correlates .50 with both first-semester scholarship (as measured by grades) and first-year scholarship. This agrees with the central tendency of similar predictions by means of a standard intelligence test.

Data on the reliability of this test are given in Table 41.

TABLE 41

DATA ON THE RELIABILITY OF THE IOWA COMPREHENSION TEST

<i>r</i>	<i>N</i>	S.D.	P.E. score	P.E. $\bar{x}.1$	P.E. score	P.E. $\bar{x}.1$	NATURE OF GROUP
					S.D.	S.D.	
.88	100	6.6	1.5	1.5	.23	.23	Grade XII

Norms. Standard scores for the Iowa Comprehension Test, Forms D-1 and D-2, are given in Table 42. They are based on April testing of high school seniors in the state of Iowa.

TABLE 42

NORMS FOR THE IOWA COMPREHENSION TEST

FORM	YEAR GIVEN	NO. CASES	NUMBER OF HIGH SCHOOLS	HIGHEST MEAN SCORE	LOWEST MEAN SCORE	MEAN (all schools)	S.D. (all schools)
D-1	1924	1854	25	24.7	18.7	21.3	6.7
D-2	1925	1869	21	27.2	15.8	25.1	7.5

Iowa High School Content Examination

Description of the test. The Iowa High School Content Examination is designed to measure high school achievement in four major fields; viz., English and literature, mathematics, science, and social studies. The four parts of the test covering these divisions are printed in a 16-page folder. The items, 400 in number, are of the 5-response type. The pupil places on a dotted line the number of the response deemed correct. The total working time of the test is 80 minutes.¹

Administration and scoring. Complete directions for taking the Iowa High School Content Examination are printed on the test blank. Scoring is completely objective, and is expedited by the use of narrow cardboard strips on which the correct answers have been accurately spaced.

Interpretation and utilization of results. The Iowa High School Content Examination has been used in several surveys of high school instruction; viz., in Iowa, Arizona, Mississippi, North Carolina, etc. It affords a good measure of general achievement. It is of value also in scholastic guidance of graduating pupils, for it indicates what information the student has actually absorbed throughout high school. At the University of Iowa the correlation between scores in the Iowa High School Content Examination and first-semester grades is .50. Since the correlation between the Iowa Comprehension Test and the Iowa High School Content Examination is only moderately high (.68), prediction of college success is improved by pooling the two series.

Validity and reliability. The Iowa High School Content Examination was confined to the four principal branches of high school instruction because agreement could be reached

¹ A recent abridgment, designated Form A-1, contains 250 items and requires 60 minutes' working time. Its reliability equals that of the longer forms.

on significant items in those subjects. Instruction in other fields — e.g., foreign language or commercial — varies so greatly that a single survey test would result in grave injustices. While similar criticism applies to some items in the content examination, such as those based on a knowledge of elementary chemistry, the test is so comprehensive that slight variations in one section tend to be outweighed or compensated by the test as a whole. The reliability data in Table 43 are based on a sampling of 247 Arizona high school seniors.¹

TABLE 43

DATA ON THE RELIABILITY OF THE IOWA HIGH SCHOOL
CONTENT EXAMINATION, FORM A

SECTION OF TEST	r	S.D.	P.E.-score	P.E. α .1	P.E.-score S.D.	P.E. α .1 S.D.
1. English93	17.5	3.1	3.1	.18	.18
2. Mathematics .	.93	18.4	2.4	2.4	.18	.18
3. Science83	11.0	3.1	2.8	.28	.25
4. History89	17.6	3.9	3.7	.22	.21
Total Score95	46.6	7.0	7.0	.15	.15

Norms. Decile norms for each of the four parts of the Iowa High School Content Examination, Form A, are given in the Manual of Directions. They are based on the testing of 1550 seniors in Iowa high schools.² Percentile scores are given also for university freshmen. Additional norms are given in Table 44. They are based on April testing of seniors in Iowa high schools.

¹ Computed by Dr. C. L. Huffaker, University of Arizona.

² See Ruch, G. M., "A Mental-Educational Survey of 1550 Iowa High School Seniors." *University of Iowa Studies in Education*, Vol. II, No. 5 (December 1, 1923); 29 pages.

TABLE 44

NORMS FOR THE IOWA HIGH SCHOOL CONTENT EXAMINATION

I. *Total Scores*

FORM	YEAR GIVEN	NO. CASES	NUMBER OF HIGH SCHOOLS	HIGHEST MEAN SCORE	LOWEST MEAN SCORE	MEAN (all schools)	S.D. (all schools)
A	1924	1827	25	185.0	139.8	160.5	52.0
B	1925	1890	21	196.8	134.5	170.5	52.4

II. *Part Scores*

SECTION OF TEST	FORM	YEAR GIVEN	NO. CASES	MEAN	S.D.
1. English	A	1923	1550	51.3	17.5
	B	1925	1890	50.1	18.5
2. Mathematics	A	1923	1550	31.6	13.4
	B	1925	1890	30.5	15.8
3. Science	A	1923	1550	28.9	11.0
	B	1925	1890	36.0	14.9
4. History	A	1923	1550	48.6	17.6
	B	1925	1890	55.1	16.2

Iowa Placement Examinations

Purpose of the examinations. The Iowa Placement Examinations are designed primarily for entering college students, but are being utilized increasingly by high schools desirous of measuring aptitude and training in specific subjects. The various tests in the placement battery have been described in the chapters devoted to high school subjects (see Chapters V-VIII). The total series comprises the following examinations:

- Chemistry Aptitude, CA-1, Revised, Forms A and B
- Chemistry Training, CT-1, Revised, Forms A and B
- English Aptitude, EA-1, Revised, Forms A and B
- English Training, ET-1, Revised, Forms A and B
- Foreign Language Aptitude, FA-1, Revised, Forms A and B
- French Training, FT-1, Revised, Forms A and B
- Spanish Training, ST-1, Revised, Forms A and B
- Mathematics Aptitude, MA-1, Revised, Forms A and B
- Mathematics Training, MT-1, Revised, Forms A and B
- Physics Aptitude, PA-1, Revised, Forms A and B
- Physics Training, PT-1, Revised, Forms A and B

What the examinations measure. Each aptitude examination measures skills essential to the work in a single field, such as mathematics. These skills are largely innate, but include the knowledge gained by earlier educational training more or less common to all students taking the test. The aptitude examinations might be considered intelligence tests of a new type, in that the items center about abilities in a single subject.

Each training examination is an educational achievement test standardized on the basis of the performance of college students, chiefly first-year students. Thus the Chemistry Training Examination is given to entering students who have had high school chemistry, in order to divide them into fast, average, and slow sections. By employing this test, college instructors are able to evaluate not only the chemical knowledge possessed by individual students, but, in time, the effectiveness of chemistry instruction in various high schools. Similarly each college department is furnished a check on the mental-educational equipment of its students in relation to a particular field of study.

Administration and scoring. The Iowa Placement Examinations can be given by any teacher without previous

practice. Each examination requires one class-hour (40 to 43 minutes' working time). Scoring is done by means of keys, and for most tests is completely objective. It is advantageous to have the scoring of the French, Spanish, and Mathematics Training Examinations supervised by a teacher in the department conducting the test.

Interpretation and utilization of results. High schools make use of the results of the Iowa Placement Examinations in the following ways:

- (1) As a basis for vocational and scholastic counseling.
- (2) For surveying the mental and educational products of the school.
- (3) For evaluating the work done by different departments in the light of standards set up for beginning college students.

In other words, they serve many of the purposes of intelligence and achievement examinations. However, their usefulness is restricted to Grades XI and XII, and primarily to guidance of pupils preparing to enter a college or technical school.

Validity and reliability. Material for the Iowa Placement Examinations was selected in accordance with a number of criteria, such as judgments of experienced college instructors, recommendations of authoritative committees, results of curriculum studies, etc. In the main, however, validity of this series will rest upon (a) the reliability of the measurement and (b) the ability of the examination to furnish real measures of the amount and character of individual differences in the mental-educational background of entering college students; i.e., of high school and preparatory school graduates. The reliability of the examinations is given in Table 45. All computations are based on a single sample of students at the University of Iowa.

TABLE 45

DATA ON THE RELIABILITY OF THE IOWA PLACEMENT EXAMINATIONS,
REVISED

NAME OF TEST	<i>r</i>	N	S.D.	P.E. score	P.E. α .1	P.E. score S.D.	P.E. α .1 S.D.
CA-1	.88	100	17.5	4.0	3.8	.23	.21
CT-1	.92	77	28.0	5.3	5.1	.19	.18
EA-1	.82	100	9.2	2.6	2.4	.29	.26
ET-1	.90	100	34.3	7.3	6.9	.21	.20
FA-1	.97	100	26.7	3.1	3.1	.11	.11
FT-1	.93	100	28.1	5.0	4.8	.18	.17
ST-1	.82	100	16.7	4.8	4.3	.29	.26
MA-1	.82	63	8.1	2.3	2.1	.29	.26
MT-1	.88	100	10.4	2.4	2.3	.23	.22
PT-1	.85	100	24.4	6.4	5.9	.26	.24

Prediction of academic success in college. Prediction of general academic success in first-year college work has been most often attempted through intelligence test scores. The correlations usually run from .30 to .60, the latter having been reached at Columbia University by means of the three parts of the Thorndike Intelligence Examination.¹ Similarly, an average correlation of .60 between test scores and first-semester grades is obtained at Iowa by combining the Iowa High School Content Examination (short form), the Iowa Comprehension Test, and the two English examinations of the placement series into a single battery. A composite of a number of Iowa Placement Examinations has proved effective as a basis of such predictions.²

¹ See Wood, Ben D., *Measurement in Higher Education* (World Book Company, 1923; 334 pages).

² See Stoddard, George D., "Iowa Placement Examinations." *University of Iowa Studies in Education*, Vol. III, No. 2 (August 15, 1925); 103 pages.

Prediction of success in specific subjects in college. Iowa Placement Examinations, given near the end of the senior year in high school or during the first week of college, can be used to predict the degree of success likely to be attained in specific subjects. This is an important service, for a student's deficiency is not always general, and comparative weaknesses in certain fields, when known, can be bolstered up by special instruction. A great number of correlations have been computed, and their central tendency is given in Table 46.

TABLE 46

CORRELATIONS BETWEEN IOWA PLACEMENT EXAMINATION SCORES AND CORRESPONDING FIRST-SEMESTER COLLEGE GRADES IN A SPECIFIC SUBJECT (APPROXIMATE CENTRAL TENDENCIES)

SUBJECT	1 SERIES (40 minutes)	2 SERIES (80 minutes)
Chemistry50	.60
English55	.60
French60	.65
Mathematics55	.60
Physics50	.55

Norms. Extensive norms based on college freshmen are available for all tests and sub-tests in the Iowa Placement series. Percentile scores for each examination are given in the Manual of Directions, and additional data may be secured from the Extension Division, University of Iowa. It is recommended that high schools utilize the norms based on college freshmen, inasmuch as the rank thus obtained places the high school graduate in his college-ability level. An abridged set of norms is given in Table 47.

TABLE 47

NORMS FOR THE IOWA PLACEMENT EXAMINATIONS, REVISED SERIES A,
BASED ON BEGINNING COLLEGE STUDENTS

EXAMINATION	NUMBER	MEAN	MEDIAN	UPPER QUARTILE	LOWER QUARTILE	STANDARD DEVIATION
CA-1	2032	58.6	59.9	73.8	43.7	20.2
CT-1	1140	73.0	70.8	97.2	46.5	34.9
EA-1	4149	40.3	39.6	46.5	32.7	9.8
ET-1	3784	92.0	90.9	120.2	65.0	40.1
FA-1	1171	75.0	74.6	95.2	54.4	27.0
FT-1	677	60.2	57.2	74.5	40.9	25.2
ST-1	606	42.1	40.1	51.0	30.7	15.8
MA-1	3104	29.1	29.2	37.2	20.6	11.5
MT-1	3780	35.6	35.5	45.0	26.0	12.9
PA-1	1888	128.1	132.0	144.6	115.1	22.4
PT-1	1275	69.5	68.8	92.3	46.2	31.8

Columbia Research Bureau Tests

Description of the tests. Six tests recently made available in the Columbia Research Bureau series cover the following subjects: plane geometry, English, French, Spanish, German, and physics. Each test is available in two equivalent forms. The tests are designed to measure achievement in the upper high school and beginning college range.

Administration and scoring. The tests can be readily given by teachers unfamiliar with testing techniques. The working time for plane geometry is 60 minutes, for physics 75 minutes, for each foreign language test 90 minutes, and for English 105 minutes. Scoring is throughout essentially objective.

Interpretation and utilization of results. The authors of the Columbia Research Bureau Tests suggest that results from the tests be employed in the following ways:

- (1) As a basis for high school grades.
- (2) As a basis for college admission credits.
- (3) As an aid in educational and vocational guidance.
- (4) As a means of improving instruction.

Norms. Extensive norms based on beginning college students are provided in the Manual of Directions for each test.

Validity and reliability. The test items of the Columbia Research Bureau battery have, for the most part, been drawn from analyses of common textbooks. Grades in Regents' examinations and college courses have also been used in evaluating test material. The reliability of these tests is unusually high, running from .90 to .97 for college freshmen.

Remedial procedures in connection with survey examinations. The chief value of survey examinations to teachers, counselors, and administrators in secondary education is clear; viz., to discover the talent most likely to succeed in higher education, and conversely, talent more likely to succeed in other fields. A great social waste, not to mention widespread individual disappointment, could be spared through a more definite and satisfactory articulation between the work of the high school and the college. The increasing use of standard tests and objective examinations has been of great value in meeting this problem. Survey examinations of the type described in this chapter may be expected to render unique service because of their acceptability as a measuring device in both high school and college.

Test Materials

Iowa Comprehension Test. By C. E. SEASHORE and others. Series D-1 and D-2, each \$1.75 per 100. Bureau of Educational Research and Service, University of Iowa, Iowa City, Iowa.

Iowa High School Content Examination. By G. M. RUCH and others. Forms B (80 minutes) and A-1 (60 minutes) each \$8.00 per 100. Bureau of Educational Research and Service, University of Iowa, Iowa City, Iowa.

Iowa Placement Examinations. The following are available:

- Chemistry Aptitude, CA-1, Revised, A and B
- Chemistry Training, CT-1, Revised, A and B
- English Aptitude, EA-1, Revised, A and B
- English Training, ET-1, Revised, A and B
- Foreign Language Aptitude, FA-1, Revised, A and B
- French Training, FT-1, Revised, A and B
- Spanish Training, ST-1, Revised, A and B
- Mathematics Aptitude, MA-1, Revised, A and B
- Mathematics Training, MT-1, Revised, A and B
- Physics Aptitude, PA-1, Revised, A and B
- Physics Training, PT-1, Revised, A and B

Each form, \$3.50 per 100, with Manual of Directions and Key. Bureau of Educational Research and Service, University of Iowa, Iowa City, Ia.

Columbia Research Bureau Tests. By BEN D. WOOD et al. Each test has Forms A and B. Prices quoted include Manual of Directions, Key, and Class Record.

- English, \$1.40 per package of 25. Specimen set, 25 cents.
- French, \$1.30 per package of 25. Specimen set, 20 cents.
- German, \$1.30 per package of 25. Specimen set, 20 cents.
- Physics, \$1.30 per package of 25. Specimen set, 25 cents.
- Plane Geometry, \$1.20 per package of 25. Specimen set, 25 cents.
- Spanish, \$1.30 per package of 25. Specimen set, 20 cents.

World Book Company, Yonkers-on-Hudson, New York.

References

- CORNOD, J., and STODDARD, GEORGE D. "Predicting Performance in Chemistry." *Journal of Chemical Education*, Vol. II, No. 8 (August, 1925).
- RUCH, G. M. "A Mental-Educational Survey of 1550 High School Seniors." *University of Iowa Studies in Education*, Vol. II, No. 5 (December 1, 1923); 29 pages.
- "College Qualifying Examinations." *School and Society*, Vol. XXI, No. 542 (May 16, 1925), pages 583-586.
- SEASHORE, C. E. "College Placement Examinations." *School and Society*, Vol. XX, No. 515 (November 8, 1924), pages 575-577.
- *The Placement Examinations as a Means for the Early Discovery and Motivation of the Future Scholar*. Nat. Research Council Bulletin; 1925.
- STODDARD, GEORGE D. "Iowa Placement Examinations." *University of Iowa Studies in Education*, Vol. III, No. 2 (August 15, 1925); 103 pages.
- "Iowa Placement Examinations." *Journal of Engineering Education*, Vol. XVI, No. 1 (September, 1925).
- "Iowa Placement Examinations, Fall of 1925, Preliminary Report." *Journal of Engineering Education*, Vol. XVI, No. 7 (March, 1926).

CHAPTER TWELVE

GENERAL INTELLIGENCE TESTS

Introduction. Intelligence tests have established a place for themselves in the technique of teaching in secondary education. Although individual intelligence tests were originally designed primarily for children, and specifically for the purpose of determining mental deficiency, the tests have now been refined and revised to such an extent that the teacher is afforded a real discrimination among pupils at all levels. Some of the uses of these tests have been indicated in Chapter II, and only one or two additional points need be added here.

- (1) Where accurate measurement is desired an *individual* intelligence test must be given; e.g., the Stanford Revision or the Herring Revision of the Binet-Simon Scale.
- (2) Administration of the revisions of the Binet Scale requires the services of an expert; at least twenty to fifty practice tests under supervision are necessary for attaining a desirable degree of proficiency.
- (3) Most of the needs in high school intelligence testing can be met by the group tests, since cases of marked mental deficiency are not ordinarily encountered.

It is recommended that the teacher or principal select a suitable group test of intelligence to be given to all students; it is desirable to have such testing done as a school project. Permanent records should be kept and made available to all instructors. They should be made available to the students only when it appears that such knowledge will be of real help. Special cases of deficiency or of extreme capacity should be retested by means of the Binet Scale. This can be done by the research department of the public school

system or by securing the services of university graduate students, psychological clinics, or departments of psychology.

The historical, critical, and interpretive literature on the nature of intelligence and the utilization of the results of intelligence tests is so extensive and so readily accessible to the teacher that no attempt will be made here to dwell upon this field. Thus the teacher or principal who is interested in this problem can get from the following books a very complete understanding of the value of intelligence tests in high schools:

1. TERMAN, L. M., *The Measurement of Intelligence*.
2. TERMAN, L. M., *The Intelligence of School Children*.
3. PINTNER, RUDOLF, *Intelligence Testing*.
4. *The Twenty-first Yearbook*, Parts I and II, of the National Society for the Study of Education: "Intelligence Tests."
5. BOEK, W. F., *The Intelligence of High School Seniors*.
6. PROCTOR, W. M., *Psychological Tests and Guidance of High School Pupils*.
7. DICKSON, V. E., *Mental Tests and the Classroom Teacher*.
8. PETERSON, JOSEPH, *Early Conceptions and Tests of Intelligence*.

Every one of the books listed above offers a different point of view, and in the aggregate they constitute an adequate library of intelligence-test information. Teachers interested in the critical aspect of intelligence testing should add to this list *Determinism in Education*, by W. C. Bagley. Needless to say, the periodical literature abounds with articles on every phase of intelligence testing.

There is, in fact, some danger that the teacher may be confused or misled by the wealth of writing in this field. Hence this chapter will be confined to specific information about intelligence tests in the high school range, organized as follows:

- (1) Individual intelligence tests.
- (2) Comparative data on group intelligence tests.
- (3) Reliabilities of intelligence tests.
- (4) Intercorrelations among intelligence tests.

- (5) Prediction of scholastic grades by means of intelligence tests.
- (6) Ways in which the high school teacher may use the results of intelligence tests.
- (7) Limitations of group testing and of intelligence testing in general.

Individual intelligence tests. The two individual tests best suited to the needs of high school teachers are the Stanford Revision and Extension of the Binet-Simon Intelligence Scale and the Herring Revision of the Binet-Simon Tests. Satisfactory administration of either of these tests requires much more than the purchase of the manuals and subsequent adherence to "rules for giving and scoring." Clinical psychologists usually insist that only persons thoroughly trained in the psychological clinic can obtain reliable results. This extreme view is scarcely tenable, since thousands of teachers throughout the country have, after a period of training, shown themselves capable of obtaining demonstrably accurate results. Mastery of the technique of individual testing depends primarily on two characteristics of the examiner: (1) the capacity to secure the complete attention and coöperation of the child — i.e., to get in *rapport* — and (2) the capacity to adhere to a rigorous procedure — i.e., to be truly scientific in method. These, after all, are characteristics of good teachers.

Training in giving the tests may be secured in the clinic or the university, or through individual study and practice. Terman recommends an apprenticeship of twenty to thirty tests in addition to psychological training. A good practice is the testing of children of known IQ; close agreement must be reached before one can be considered proficient enough to make his results a matter of record. In large school systems the high school teacher is rarely called upon to give individual

intelligence tests; a highly trained personnel does the work, and IQ's of all pupils are available as part of their scholastic record. The same personnel sometimes offers a training course for teachers interested in Binet testing. In smaller systems it is advantageous to have only a few teachers give the tests; they can be better trained, and the service can be more efficiently rendered.

The Stanford Revision and the Herring Revision may be considered equivalent tests; they cover the same range, the contents are similar, and both lead to a mental age and an intelligence quotient. The Herring Revision is a point-scale, with mental-age equivalents given in tables. The training required is about the same for both. Herring states that his revision will correlate .99 with the Stanford Revision for unselected age groups, and that an examiner has attained sufficient proficiency "when he can maintain a correlation of .97" between the Herring and Stanford revisions.

Comparative data on group intelligence tests in the high school range. Useful information is given in Table 48. Principals or superintendents contemplating extensive intelligence testing should be guided not only by the technical evaluation of the tests, but by such practical considerations as economy of time and money in giving and scoring, and the definite utility of the results obtained.

Reliability of intelligence tests. Coefficients of reliability for various group intelligence tests are given in Table 49. These coefficients are comparable to those obtained for the better educational achievement tests. This is to be expected, since intelligence tests usually sample a great number of skills. High reliability is not, however, a guarantee of high validity; i.e., that we really have an adequate test of *mental ability*.

TABLE 48. COMPARATIVE DATA ON GROUP

NAME OF TEST	AUTHOR	No. FORMS	No. PARTS	No. ITEMS	WORKING TIME
Army Alpha	Committee	5	8	212	45 minutes
Brown University	S. S. Colvin	1	10	150	60 minutes
Dearborn	W. F. Dearborn	1	7	II: 171	65 minutes
Haggerty, Delta 2	M. E. Haggerty	1	6	176	35 minutes
McCall (Multi-Mental)	W. A. McCall	1	1	100	25 minutes
Miller	W. S. Miller	2	3	120	30 minutes
Otis { Advanced Self-Administering	A. S. Otis	2 2	10 1	230 75	42 minutes 30 minutes
Pressey { Classification Verification	S. L. Pressey	1 1	1 1	96 96	16 minutes 16 minutes
Terman	L. M. Terman	2	10	185	35 minutes
Thorndike	E. L. Thorndike	Numer- ous	I: 13 II: 8 III: 3	I: 136	Practice Form: 10 minutes I: 30 minutes II: 60 minutes III: 60 minutes
Thurstone	L. L. Thurstone	1	6	168	30 minutes

INTELLIGENCE TESTS FOR HIGH SCHOOLS

FORE-EXERCISE	VALIDATION	NORMS	RANGE OF TALENT	REMARKS
Yes	See <i>Memoirs of National Academy of Science</i> , XV (1921)	Extensive	Gr. VII, up	Of historical interest
Yes, 10 minutes	See Colvin in <i>School and Soc.</i> , X	Grade	Gr. XII, up	
Yes		Age & Grade	Gr. IV-XII	Largely pictorial
Yes		Age & Grade	Gr. III-IX	
Yes	See McCall in <i>Teachers College Record</i> , XXVII, Oct. 1925 & Jan. 1926	Age, Grade and T	Gr. II-XII	Primarily for Gr. III-VIII. The most recent intelligence test
Yes	See <i>21st Yearbook of N.S.S.E.</i>	Extensive	Gr. VII-XII	Limited range of skills
Yes	See Otis in <i>Jour. Educ. Psych.</i> , IX	Ages 8-18 Age & Grade	H. S. & Col. Gr. IX-XII	Complete data in Manual of Directions
Yes		Age & Grade	Gr. VII-XII	Both tests should be given
Yes		Medians		
Yes	See World Book Co Booklet	Extensive Age & Grade	Gr. VII-XII	The most reliable group intelligence test for H. S.
Yes, 10 minutes	See Wood: <i>Measurement in Higher Education</i>	Extensive	Gr. XII, up	Part I only recommended for H. S. seniors
	See Thurstone in <i>Jour. Educ. Psych.</i> , X	H. S. and College	H. S. and College	Especially for college freshmen

TABLE 49

RELIABILITY OF INTELLIGENCE TESTS

NAME OF TEST	<i>r</i>	NO. CASES	AGE RANGE	GRADE RANGE	SOURCE
Haggerty, Delta 2	.88	40	15-6 to 16-5 Median C. A.:	— H. S.	Haggerty
Miller	.91	109	16-3	Sophomores	Miller
Otis (Self-Administering)	.92	128	—	IX-XII	Otis
Stanford-Binet (IQ)	.93	428	3 to 15	—	Terman
Stanford-Binet	.95	114	(Adults)	—	Rugg

Intelligence-test intercorrelations. The extent to which different intelligence tests really measure the same capacities is indicated in Table 50. These coefficients of correlation can be interpreted only in the light of the age and grade range of the pupils tested. Tests which (for the same range of talent) correlate as highly as do two forms of a single test (see Table 49) may be used interchangeably.

TABLE 50

INTELLIGENCE-TEST INTERCORRELATIONS

TESTS	<i>r</i>	NO. CASES	GRADE RANGE	SOURCE
Army Alpha <i>vs.</i> Haggerty, Delta 2	.76	55	IX	Haggerty
Terman <i>vs.</i> Haggerty, Delta 2	.75	55	IX	Haggerty
Otis <i>vs.</i> Haggerty, Delta 2	.73	55	IX	Haggerty
Army Alpha (IQ) <i>vs.</i> Stanford-Binet (IQ)	.74	116	IX-XII	Proctor
Miller <i>vs.</i> Haggerty, Delta 2	.78	55	IX	Miller
Miller <i>vs.</i> Terman (A)	.75	55	IX	Miller
Miller <i>vs.</i> Army Alpha (8)	.76	55	IX	Miller
Miller <i>vs.</i> Otis	.73	55	IX	Miller

Further data on the relationships among various tests of intelligence are given in Tables 51, 52, and 53. It is clear that, for a restricted range of talent, mental ages or IQ's derived from different group intelligence tests are only roughly comparable.

TABLE 51

CORRELATIONS OF FIVE INTELLIGENCE TESTS FOR 64 PUPILS OF HIGH SCHOOL AGE ACCORDING TO JORDAN'S RESULTS¹

Binet Mental Ages with:		
	Otis Group Test66 ± .047
	Army Alpha69 ± .044
	Miller Group Test53 ± .060
	Terman Group Test68 ± .045
Composite of Alpha, Otis, Miller, and Terman with:		
	Otis Group Test93 ± .012
	Army Alpha91 ± .015
	Miller Group Test90 ± .016
	Terman Group Test91 ± .015
Teachers' Estimates of Intelligence with:		
	Otis Group Test73 ± .039
	Army Alpha61 ± .052
	Miller Group Test68 ± .045
	Terman Group Test66 ± .047

¹ Jordan, A. M., "The Validation of Intelligence Tests." *Journal of Educational Psychology*, Vol. XIV (1923), pages 348-366, 414-428.

TABLE 52. CORRELATION OF INTELLIGENCE TESTS IN HIGH SCHOOL GRADES ACCORDING TO ROOT'S INVESTIGATION¹

TESTS	GRADES		
	9	10	11-12
Binet with :			
Terman Group Test, A35(22) ²	.67(25)	.53(37)
Otis Adv. Exam., A72(22)	.55(25)	.44(37)
Haggerty, Delta 244(22)		
Mentimeters43(22)	.68(25)	.54(36)
Dearborn, Series II47(22)		
Terman Group Test, A, with :			
Otis Adv. Exam., A73(21)	.87(25)	.72(35)
Haggerty, Delta 285(21)		
Mentimeters60(20)	.79(22)	.63(34)

TABLE 53. AVERAGE CORRELATIONS OF FOURTEEN GROUP INTELLIGENCE TESTS WITH EACH OF THE OTHER THIRTEEN TESTS ACCORDING TO FRANZEN³

TEST	AVERAGE CORRELATION
Terman Group Test75
National A74
Haggerty73
Illinois72
Otis71
Mentimeter66
Pressey Survey65
National B62
Thorndike Reading59
Dearborn — 158
Pressey Cross-Outs56
Dearborn — 255
Wylie Opposites53
Myers46

The number of cases is 57 throughout.

¹ Root, W. T., "Correlations between Binet Tests and Group Tests." *Journal of Educational Psychology*, Vol. XIII (1922), pages 286-292.

² The numbers of cases used are given in parentheses.

³ Franzen, R., "Attempts at Test Validation." *Journal of Educational Research*, Vol. VI (1922), pages 145-158.

Prediction of high school marks. Prediction of school marks for pooled subjects in high school is given in Table 54. The correlations are comparable only for the same grade range.

TABLE 54¹

PREDICTION OF SCHOOL MARKS IN POOLED HIGH SCHOOL SUBJECTS

NAME OF TEST		NUMBER OF CASES	GRADE RANGE	SOURCE
Haggerty, Delta 256	55	IX	Haggerty
Army Alpha34	494	IX-XII	Proctor
Army Alpha41	480	IX-XII	Proctor
Miller56	55	IX	Miller
Haggerty, Delta 250	55	IX	Miller
Terman (A)59	55	IX	Miller
Army Alpha (S)56	55	IX	Miller
Stanford-Binet (IQ) . .	.45	111	IX-XII	Terman

Intelligence test scores have been frequently correlated with grades in individual high school subjects. The relationship between measures of intelligence and English is usually the highest found, but it is not close. For class organization and individual diagnosis in a single subject, a test of achievement can best be supplemented by an aptitude or prognosis test for the particular subject.

Utilization of results of intelligence testing. A knowledge of the mental ability of each high school pupil is valuable in a number of ways. The various references previously mentioned, as well as the manuals published with the tests, indicate in some detail just what use can be made of the scores, mental ages, and IQ's obtained. These uses may be listed as follows:

¹ Data for Tables 49, 50, and 54 were obtained from recent literature, as indicated. See References at end of this chapter.

1. For classification of students in a subject. In large high schools it is desirable to have fast and slow sections. Intelligence ratings give important information here, but they should not be made the sole criterion of selection.
2. In aiding students to select subjects and courses. Skillful counseling necessitates reliable information about the pupil's mental ability.
3. In vocational guidance. Teachers are likely to place too much dependence upon intelligence test results in meeting the difficult problems of guidance. For any one level of intelligence a great number of life activities are appropriate and desirable; intelligence simply marks off rough dividing lines between careers which increasingly call upon the power of abstraction. Definite choice of a career must take into account non-intellectual aptitudes, individual interests, emotional traits, and economic and social needs. Intelligence tests, for example, have but restricted application in trade education and guidance. Thus special tests have been devised to meet industrial needs (see Chapter X).
4. To indicate the pupil's capacity for higher education. Numerous studies have shown that many of the maladjustments and many failures among college students are due to inadequate mental ability. The student whose record shows a low IQ and poor work in high school is not likely to succeed in college. However, in making decisions of such importance, the teacher or counselor should have at his command the IQ from an *individual* test, and a rather complete account of the pupil's outstanding characteristics and previous scholastic achievements. The intelligence test is of prime importance, on the other hand, in the discovery of gifted pupils, for whom no expenditure of educational effort and guidance is too great. It is not rare for such children to pass unnoticed through a school system, until brought to light through the findings of a mental test. Is it not true that in many in-

stances these exceptional children are given, through the medium of a mental test, their first genuine opportunity to demonstrate intellectual superiority?

Limitations of intelligence testing. The principal limitations have already been touched upon, and it is clear that they center about the *abuse* of the knowledge represented by a point-score, mental age, or IQ. Every such rating represents a complex interweaving of innate and acquired abilities, difficult to define and interpret. Interpretation, no less than the administration of intelligence tests, requires definite training on the part of the teacher or counselor. In the group test, administration and scoring are reduced to a simple and rapid procedure, but the difficulty of interpretation is increased. A sane use of intelligence ratings demands that they be treated always in the light of the whole body of available pupil-information.

Test Materials

Army Group Intelligence Examination, Alpha. \$8.00 per 100 booklets; Manual of Directions, 75 cents; stencils, \$1.25. Specimen set, 80 cents. Bureau of Educational Measurements, Kansas State Teachers' College, Emporia, Kansas. Form 8, \$6.75 per 100. C. H. Stoelting Company, Chicago, Illinois.

Brown University Psychological Examination. By S. S. COLVIN. Series II. J. B. Lippincott Company, Philadelphia.

Dearborn Group Intelligence Tests. By W. F. DEARBORN. Series II, C and D, \$4.50 per 100. Manual of Directions, 25 cents. J. B. Lippincott Company, Philadelphia.

Haggerty Intelligence Examination, Delta 2. By M. E. HAGGERTY. \$1.10 per package of 25, with Key and Class Record. Manual of Directions (58 pages), 25 cents. Specimen set, 50 cents. World Book Company, Yonkers-on-Hudson, New York.

Herring Revision of the Binet-Simon Tests. By JOHN P. HERRING. Examination Manual: Form A (56 pages), \$1.00. Individual Record Card, \$1.00 per package of 25. World Book Company, Yonkers-on-Hudson, New York.

Multi-Mental Scale. By W. A. McCALL. One copy needed for each pupil; \$1.00 per 100. Manual of Directions and Scoring Stencil, 15 cents.

Bureau of Publications, Teachers College, Columbia University, New York.

Miller Mental Ability Test. By W. S. MILLER. Forms A and B, each 80 cents per package of 25, including Key, Age-Grade-Score Sheet, and Percentile Graph. Manual of Directions (24 pages), 15 cents. Specimen set, 25 cents. World Book Company, Yonkers-on-Hudson, New York.

Otis Group Intelligence Scale, Advanced Examination. By A. S. OTIS. Forms A and B, each \$1.25 per package of 25, with Key and Record Sheet. Manual of Directions (80 pages), 30 cents. Specimen set, 50 cents. World Book Company, Yonkers-on-Hudson, New York.

Otis Self-Administering Tests of Mental Ability, Higher Examination. By A. S. OTIS. Forms A and B, each 80 cents per package of 25, with Manual of Directions and Key, Interpretation Chart and Percentile Graph, and Class Record. Specimen set, 30 cents. World Book Company, Yonkers-on-Hudson, New York.

Pressey Senior Classification Test. By S. L. PRESSEY. \$1.25 per 100. Specimen set, 10 cents. Public School Publishing Company, Bloomington, Illinois.

Pressey Senior Verification Test. By S. L. PRESSEY. \$1.25 per 100. Specimen set, 10 cents. Public School Publishing Company, Bloomington, Illinois.

Stanford Revision of the Binet-Simon Intelligence Scale. By L. M. TERMAN. Complete material, \$11.00. C. H. Stoelting Company, Chicago. Package of Test materials, including Record Booklet, \$1.00; additional Record Booklets, \$2 per package of 25. (Other equipment can be made by examiner.) Houghton Mifflin Company, Boston.

Terman Group Test of Mental Ability. By L. M. TERMAN. Forms A and B, each \$1.20 per package of 25, with Key, Record Sheet, and Manual of Directions. Specimen set, 15 cents. World Book Company, Yonkers-on-Hudson, New York.

Thorndike Intelligence Examination for College Entrance. By E. L. THORNDIKE. Price varies. Bureau of Publications, Teachers College, Columbia University, New York.

Thurstone Psychological Examination for College Freshmen and High School Seniors. By L. L. THURSTONE. \$15.00 per 100. C. H. Stoelting Company, Chicago.

References

BAGLEY, W. C. *Determinism in Education.* Warwick & York, Inc., Baltimore; 1925. 194 pages.

BALDWIN, B. T., and STECKER, L. I. "Mental Growth Curve of Normal and Superior Children." *University of Iowa Studies in Child Welfare*, Vol. II, No. 1 (1922); 61 pages.

BOOK, W. F. *The Intelligence of High School Seniors*. The Macmillan Company, New York; 1922. 371 pages.

BRIGHAM, CARL C. *A Study of American Intelligence*. Princeton University Press, Princeton, New Jersey; 1923. 210 pages.

DEARBORN, W. F., and LINCOLN, E. A. "How the Dearborn Intelligence Examination Standards Were Obtained." *Journal of Educational Psychology*, Vol. XIII (May, 1922), pages 295-297.

DICKSON, VIRGIL E. *Mental Tests and the Classroom Teacher*. World Book Company, Yonkers-on-Hudson, New York; 1925. 231 pages.

FEINGOLD, G. A. "Correlation between Intelligence and Scholarship." *School Review*, Vol. XXXII (June, 1924), pages 455-467.

HAGGERTY, M. E. "Intelligence Examination: Delta 2." *Journal of Educational Psychology*, Vol. XIV (May, 1923), pages 257-276.

"Intelligence Tests and Their Uses." *The Twenty-first Yearbook of the National Society for the Study of Education*, Parts I and II (1922); 270 pages.

MCCALL, W. A., and his Students. "The Multi-Mental Scale." *Teachers College Record*, Vol. XXVII, No. 2 (October, 1925), pages 109-120.

— "Construction of the Multi-Mental Scale." *Teachers College Record*, Vol. XXVII, No. 5 (January, 1926), pages 394-399.

MACPHAIL, A. H. *The Intelligence of College Students*. Warwick & York, Inc., Baltimore; 1924. 176 pages.

ODELL, C. W. *Conservation of Intelligence in Illinois High Schools*. Bulletin No. 22 (1925); 55 pages. Bureau of Educational Research, University of Illinois, Urbana, Illinois.

OTIS, ARTHUR S. "An Absolute Point Scale for the Measurement of Intelligence." *Journal of Educational Psychology*, Vol. IX (May, 1918), pages 239-261; (June, 1918), pages 333-348.

— and KNOLLIN, H. E. "The Reliability of the Binet Scale and of Pedagogical Scales." *Journal of Educational Research*, Vol. IV (September, 1921), pages 121-142.

PETERSON, JOSEPH. *Early Conceptions and Tests of Intelligence*. World Book Company, Yonkers-on-Hudson, New York; 1925. 320 pages.

PINTNER, RUDOLF. *Intelligence Testing*. Henry Holt & Co., New York; 1923. 406 pages.

— and MARSHALL, H. "A Combined Mental-Educational Survey." *Journal of Educational Psychology*, Vol. XII (January, 1921), pages 32-43; (February, 1921), pages 82-91.

PROCTOR, W. M. "Psychological Tests and Guidance of High School Pupils." *Journal of Educational Research Monograph*, No. 1 (October, 1923); 125 pages.

TERMAN, L. M. *The Intelligence of School Children*. Houghton Mifflin Company, Boston; 1919. 317 pages.

— *The Measurement of Intelligence*. Houghton Mifflin Company, Boston; 1916. 362 pages.

TERMAN, L. M., et al. *Intelligence Tests and School Reorganization*. World Book Company, Yonkers-on-Hudson, New York; 1925. 111 pages.

THORNDIKE, E. L. "The Measurement of Intelligence." *Psychological Review*, Vol. XXXI (May, 1924), pages 219-252.

VITELES, M. S. "Psychological Tests in Guidance: Their Use and Abuse." *School and Society*, Vol. XXII, No. 560 (September 19, 1925), pages 350-356.

WOOD, BEN D. *Measurement in Higher Education*. World Book Company, Yonkers-on-Hudson, New York; 1923. 337 pages.

YERKES, R. M. (Editor). "Psychological Examining in the United States Army." *Memoirs of the National Academy of Sciences*, Vol. XV (1921); 890 pages.

YOAKUM, C. S., and YERKES, R. M. *Army Mental Tests*. Henry Holt & Co., New York; 1920.

CHAPTER THIRTEEN

JUNIOR HIGH SCHOOL TESTS

Introduction. The problem of measurement in the junior high school merits more attention than has been paid to it by test builders. A survey of standard tests available for this period leads to the conclusion that few tests are constructed in accordance with the 6-3-3 plan of secondary school organization. Hence the testing needs of junior high school teachers must be largely met, for the present, by standardized tests designed primarily for intermediate or senior high school grades. In either case the range of talent of the pupils tested is likely to fall outside the most effective measuring range of the test or scale.

The scope of measurement in the junior high school. Recent curriculum studies point out the gradual emergence of a body of approved subjects appropriate to the junior high school. Table 55 was assembled from data reported by J. M. Glass.¹ The means tabulated in this table are computed on the basis of positive cases only; i.e., they represent the average number of minutes devoted to a certain subject in those junior high school courses of study which included the subject at all.

The report of the committee which in 1923 surveyed the junior high schools of New York City contains many suggestive courses of study. Table 56 presents the minimum essentials recommended for pupils in classes making normal progress.

Description of tests for junior high school grades. It is somewhat beyond the purpose of the present volume to give a detailed evaluation of the tests and scales which fall into the extensive area implied in the courses of study for

¹ Glass, J. M., *Curriculum Practices in the Junior High School and Grades 5 and 6* (University of Chicago, 1924; 181 pages).

TABLE 55
(Adapted from Glass)

NUMBER OF MINUTES PER WEEK ALLOTTED FOR REQUIRED SUBJECTS.
COMPOSITE RESULTS FROM 14 CITIES

NAME OF SUBJECT	AVERAGE MINUTES PER WEEK		
	Grade VII	Grade VIII	Grade IX
English Composition	82	80	89
English Grammar	83	63	49
English Literature and Reading	117	100	118
Spelling	39	31	26
Physical Education	111	115	124
Home Economics	227	177	283
Algebra	53	104	216
Arithmetic	192	165	175
Geometry	90	75	85
General Science	143	178	195
Hygiene	45	41	41
Social Studies	372	405	224
Penmanship	69	71	—
Industrial Arts	144	195	183

junior high schools. Many of the tests used will be found treated at some length in books devoted to measurement in the elementary and intermediate grades. But a number of tests commonly employed will be brought together here, with the thought that even brief descriptions and recommendations may prove helpful to the classroom teacher.

For a description of tests which may be used in both junior and senior high schools, such as algebra, English, general science, etc., the reader is referred to the chapters devoted to the various high school subjects.

I. MATHEMATICS TESTS

Buckingham Scale for Problems in Arithmetic, Division III. This scale is based upon the ability of pupils to solve verbal

TABLE 56¹MINIMUM NUMBER OF 45-MINUTE PERIODS PER WEEK,
NEW YORK JUNIOR HIGH SCHOOLS

SUBJECT	GRADE					
	7A	7B	8A	8B	9A	9B
English	8	8	7	7	5	5
Foreign Language (or an elective)	—	—	(3)	(3)	(4)	(4)
Arithmetic and Algebra	4	4	4	4	4	4
History and Civics	3	3	3	3	—	—
Community Civics and Current History	—	—	—	—	2	2
Geography	2	2	(2)	(2)	—	—
Elementary Science (Boys)	2	2	2	2	—	—
Biology	—	—	—	—	5	5
Drawing	2	2	2	2	2	2
Shopwork (Boys)	2	2	2	2	(2)	(2)
Cooking (Girls)	2	2	2	2	2	2
Sewing (Girls)	2	2	2	2	(2)	(2)
Music	2	2	2	2	2	2
Physical Training and Hygiene	3	3	3	3	3	3
Typewriting	—	—	—	—	(3)	(3)

problems in arithmetic. It is designed for Grades VII and VIII. Use Form I for the first testing, and Form II for the second testing. The two forms are approximately equivalent.

Compass Diagnostic Tests, Form A. This series consists of twenty diagnostic tests developed through extensive experimental procedure. Ninety different arithmetic skills are measured, and each test is of sufficient reliability to enable individual diagnosis. This series of tests is by far the most comprehensive attempt at the measurement of arithmetical abilities, especially in Grades VI, VII, and VIII. Each test

¹ From *Survey of the Junior High Schools of New York (City)* (Board of Education, New York, 1924; 257 pages).

should be given upon completion of the instructional unit to which it applies. The tests cover Grades II to VIII. They are easy to score. The titles of the tests and the appropriate grades are given under Test Materials.

Courtis Standard Research Tests, Series B. These tests have been widely used and extensive norms are available. They measure the speed and accuracy of performance for addition, subtraction, multiplication, and division of whole numbers. They are designed for Grades IV to VIII, but are of little use in the junior high school.

Monroe General Surrey Scales in Arithmetic. Scale II is designed for Grades VI, VII, and VIII, and is available in three equivalent forms. The first four sub-tests of this scale are similar to the Courtis tests, but the scale also includes operations with fractions and decimals.

Monroe Standardized Reasoning Tests in Arithmetic, Test III. This test is designed for Grade VIII. It consists of fifteen problems which are scored for both the correct principle and the correct answer. The two forms are not equivalent.

Otis Arithmetic Reasoning Test. This is simply Test 5 of the Otis Group Intelligence Scale, which is described in Chapter XII. It is printed as a separate test.

Spencer Diagnostic Tests in Arithmetic, III. This test can be used for class and individual diagnosis. In addition to the number of problems correctly solved a record is kept of the kind and number of errors made. Two forms are available for Grades VII and VIII.

Stanford Arithmetic Examination. Parts 4 and 5 of the Stanford Achievement Test are published as a separate test, comprising arithmetic computation and reasoning. The test is diagnostic for these skills. Reliability coefficients for the junior high school range are given below:

AGE	ARITHMETIC COMPUTATION	ARITHMETIC REASONING
13	.85	.91
14	.83	.88
15	.76	.89

Stone Reasoning Test. The Stone Reasoning Test represents an early attempt to measure arithmetical ability. It has been widely used in city surveys. The test may be used in Grades V to VIII, but it is not very reliable and is difficult to score.

Woody Arithmetic Scales. These are published in two series, A and B, and can be used in Grades II to VIII. They measure achievement in arithmetic by means of four scales, one for each of the fundamental operations. Series A is longer than Series B, and is to be recommended.

Woody-McCall Mixed Fundamentals, Forms I, II, III, and IV. These scales represent a combination of all the operations in the Woody Arithmetic Scales. The forms are equivalent and may be used in Grades III to VIII. Both Woody Arithmetic tests are fairly reliable.

Wisconsin Inventory Test in Arithmetic. Eight tests are available as listed under Test Materials. They are designed to give a measure of separate arithmetic skills in order that the teacher may adjust the work of the class accordingly. They cover Grades II to VIII, but are chiefly useful in Grades III to VI.

II. ENGLISH TESTS

Briggs English Form Test. Parallel forms, Alpha and Beta, measure seven important rules of punctuation. They may be used in Grades VII-IX. The working time is 20 minutes. The reliability of the test, as computed from 38 pupils in Grades IX, was found to be .80, and the probable error of a score, 1.5.

Charters Diagnostic Language and Grammar Tests. The Language Test is designed for Grades III to VIII, and the Grammar Test for Grades VII and VIII. They measure knowledge of verbs and pronouns, and miscellaneous constructions. The reliability, as determined from 80 ninth-grade pupils, is .78, and the probable error of a score, 2.4.

Franseen Diagnostic Tests in Language. Three parts, printed on separate folders, cover respectively Pronouns, Verbs, and Varied Constructions. They are much more comprehensive than the Charters series. Scoring is completely objective. The Franseen tests are intended to be a teaching device for Grades III to VIII.

Ayres Spelling Scale. This scale embodies one of the earlier attempts at measurement. It is too easy above seventh grade. The Buckingham Extension includes 505 words in addition to the original Ayres 1000.

Iowa Dictation Exercise and Spelling Test. Form C is designed for Grades VII and VIII. It is based on the 73 per cent column of the Iowa Spelling Scale. Norms for context and list forms are available.

Stanford Dictation Exercise. This spelling test comprises Test 9 of the Stanford Achievement Test. It is based upon the studies of Ayres, Buckingham, and Thorndike. All the major words contained in the sentences dictated enter into the spelling measurement. Its reliability for unselected age groups is .90. The test is available only in the Stanford Achievement Test booklet.

Chapman-Cook Speed of Reading Test. This test is designed to measure how rapidly pupils can read with full comprehension. Norms are given for ten levels of achievement in each grade. It is designed for Grades IV to VIII.

Gray Standardized Oral Reading Check Tests. Test IV enables a rather complete analysis of individual difficulties in oral reading in Grades VI to VIII. The test is given individually. Administration and scoring are laborious.

Monroe Standardized Silent Reading Tests, Revised. Test II is designed for Grades VI to VIII. Forms 1 and 2 are approximately equivalent. The test is too brief to be of value in measuring individual pupils.

Stanford Reading Examination. This test consists of three parts of the Stanford Achievement Test, published in a separate folder; viz., Paragraph Meaning, Sentence Meaning, and Word Meaning. The test accurately diagnoses these three reading abilities. There are two equivalent forms. Carefully derived norms (age, grade, and *T*-score) are given in the Stanford Achievement Test Manual. Reliability data for the junior high school range are given below:

AGE	PARAGRAPH MEANING	SENTENCE MEANING	WORD MEANING
13	.91	.89	.96
14	.91	.92	.96
15	.82	.91	.94

Stone Narrative Reading Tests. In this test the pupil reads two short narratives. Comprehension of the subject matter and rate of reading are scored. The same sheets can be used with succeeding classes.

Thorndike Test of Word Knowledge. Each of the four forms of this test consists of a hundred words graded according to importance, selected from the 10,000 commonest words contained in Thorndike's *The Teacher's Word Book*. The test was given to thousands of school children by the Classical League of America. Norms are available for Grades IV to IX.

Additional information on reading tests and comparative reliabilities of common reading tests are given in Table 57.¹

¹ From Current, W. F., and Ruch, G. M., "Further Studies on the Reliability of Reading Tests." *Journal of Educational Psychology*, September, 1926, pages 476-481.

TABLE 57
RELIABILITIES OF SIX WELL-KNOWN READING TESTS

TEST	r	N	S.D. ₁	S.D. ₂	P.E. _{score}	P.E. _{∞,1}	P.E. _{score} S.D.	P.E. _{∞,1} S.D.	M ₁	M ₂	NATURE OF GROUP
Lippincott-Chapman	.89	154	7.3	—	1.6	1.5	.22	.21	16.2	—	Grades 4-8 (Same pupils throughout)
Curtis	.74	154	12.9	13.4	4.5	3.9	.34	.30	35.3	44.3	"
Haggerty	.88	154	27.7	19.7	6.6	6.0	.28	.25	53.3	52.3	"
Monroe (Comp.)	.76	154	3.4 ¹	3.1 ¹	1.1 ¹	0.9 ¹	.33	.29	9.8	9.9	"
Thorn-McCall	.76	154	4.5	4.4	1.5	1.3	.34	.29	21.1	19.2	"
Stanford	.93	154	40.8	42.5	7.4	7.1	.18	.17	138.8	142.1	"

¹ Due to an accident, the correlation sheets upon which the Monroe test correlations were computed were lost. The values given above for the standard deviations are in terms of "steps" as grouped in computing the correlation. It cannot be ascertained at this time what the grouping factor was. Since the data in the other columns for the Monroe test have been carried out using steps rather than actual scores, the P.E./S.D. ratios are strictly comparable with those of the other tests.

Willing Scale for Measuring Written Composition. Compositions actually written by pupils in Grades IV to VIII form the basis of this scale. The scale yields separate measures of story value and form value, the latter obtained by counting the errors in grammar. The scale has been widely used.

III. GEOGRAPHY TESTS

Buckingham-Sterenson Place Geography Tests. These tests measure the pupil's knowledge of geographical locations. Three forms of each of the two tests (the World and the United States) are available. Norms for Grades IV to VIII are given. The Buckingham-Stevenson Information Test measures knowledge of the geography of the United States, South America, and Europe.

Courtis Supervisory Geography Test. Test A measures the pupil's ability to name states and locate cities on outline maps. Test B deals with oceans and continents.

Gregory-Spencer Geography Tests. These tests are much more comprehensive than the Buckingham and Courtis tests. The important phases of geography covered are: trade routes, causal geography, place and description geography, physical and commercial geography, and map study. Three forms are available. The tests are designed for Grades VI to VIII.

Hahn-Lackey Geography Scale. This scale is similar in construction to the Ayres Spelling Scale. Questions are drawn from the scale to make a test in accordance with classroom needs.

Posey-Van Wagenen Geography Scales. Division II is designed for Grades VII and VIII. Information Scale (R) is paralleled by a Thought Scale (S). In addition to these general information scales another series (A, F, and K) deal with the various continents. The scales are difficult to

administer, and the data in Tables 58 and 59 indicate other weaknesses.

Comparative data on geography tests. Tables 58, 59, and 60 are quoted from a recent study of the reliability and validity of geography tests in common use.¹ Table 58 gives the correlation between geography tests and a criterion of pupil performance in geography derived from combining the results of 108 objective examinations. This experimental work covered nine grades in three schools.

TABLE 58
VALIDITY COEFFICIENTS OF GEOGRAPHY TESTS

TEST	<i>r</i>	P.E. <i>r</i>	<i>N</i>
Buckingham-Stevenson, Place69	.03	72
Courtis66	.03	123
Buckingham-Stevenson, Information54	.04	123
Hahn-Lackey "A"50	.04	126
Gregory-Spencer A45	.05	127
Posey-Van Wagenen, Thought R29	.05	127

Table 59 gives reliability data for the geography tests. The figures in the last two columns afford a comparable measure of reliability among the six tests.

Intercorrelations, with their probable errors, are given for geography tests in Table 60. They indicate a heterogeneity of functions now being measured by tests in geography.

¹ Eyestone, A. B., and Ruch, G. M., *Studies on Standard Tests in Geography*. Unpublished.

TABLE 59

RELIABILITY COEFFICIENTS FOR GEOGRAPHY TESTS

TEST	N	GRADES	S.D. ₁	S.D. ₂	r_{12}	P.E. score	P.E. ∞	P.E. ∞	P.E. ∞
								SD.	S.D.
Courtis Location . . .	166	V-VII	11.8 ¹	11.9 ²	.92 ³	2.3	2.2	.19	.18
Buckingham-Stevens- son Inf. P. . . .	195	V-VII	5.8 ¹	6.5 ²	.95 ⁴				
Posey-Van Wagenen Thought R. . . .	169	V-VII	9.1		.77 ³				
Gregory-Spencer . .	168	V-VII	20.2	18.6	.58	4.0	3.0	.44	.33
Hahn-Lackey	175	V-VII	5.6	5.6	.81 ⁵	5.7	5.1	.29	.26
Buckingham-Stevens- son Place	82	V-VII	44.8 ⁶		.86 ⁷	11.2	10.4	.25	.23

TABLE 60

INTERCORRELATIONS AMONG GEOGRAPHY TESTS

TEST	2	3	4	5	6
1. Courtis	.58 \pm .03	.47 \pm .04	.68 \pm .03	.64 \pm .03	.85 \pm .02
2. Buckingham-Stevenson Infor- mation-Problems, U. S. Form 1		.53 \pm .04	.69 \pm .03	.58 \pm .03	.72 \pm .02
3. Posey-Van Wagenen Thought R.			.51 \pm .04	.56 \pm .04	.52 \pm .05
4. Gregory-Spencer A				.75 \pm .02	.57 \pm .05
5. Hahn-Lackey "A" (30 items)					.72 \pm .04
6. Buckingham-Stevenson, Place Geography, I					

¹ Odd-numbered items.² Even-numbered items.³ Correlation of odds and evens.⁴ Estimated for whole form by Spearman-Brown formula.⁵ r for 30 items against 30 items (For 4 items, r is about .36; for 10 items, about .61).⁶ Average sigma for three forms.⁷ Average r for 3 reliability coefficients figured on all three forms.

IV. VOCATIONAL SUBJECTS

Ayres Measuring Scale for Handwriting, Gettysburg Edition. Specimens of children's handwriting are presented on a regular scale. Standards are available for speed and quality. Models are given for vertical, semi-slant, and slant specimens at each step of the scale.

Courtis Standard Practice Tests in Handwriting. This series consists of twenty graded lessons designed for Grades III-VIII, for use with the Ayres Scale.

Freeman Chart for Diagnosing Faults in Handwriting. This scale diagnoses (a) uniformity of slant, (b) uniformity of alignment, (c) equality of line, (d) letter formation, and (e) spacing. The test is rather long as a scale for measuring general quality. It is of great value in the discovery of pupil deficiencies.

Thorndike Handwriting Scale. For General Merit of Children's Handwriting. Fifteen qualities of handwriting are presented, with one or more specimens of handwriting for each quality. The test can be used for Grades II to VIII. Standards for quality and speed are printed on the scale.

Gates-Strang Health Knowledge Test. This is a recent test which can be used in Grades III to XII. It consists of 64 exercises designed to measure a pupil's knowledge of the phases of healthful living.

Short Scales for Measuring Habits of Good Citizenship. These consist of four scales derived from the 1919 study of Upton and Chassell on the measurement of good citizenship. The scales are equivalent and may be used up to Grade VIII.

Home Economics Information Tests. These tests are designed for girls completing the eighth grade, and appear in three test booklets. Set I contains questions on Clothing, Set II, Food, and Set III, Other Problems of the Home.

King-Clark Foods Test. This test permits the classification of pupils from Grades VI to XII with respect to knowledge of foods. It is based upon an analysis of nine common textbooks.

Murdoch Sewing Scale. This scale represents a careful attempt to measure accomplishment in sewing. It is useful in grading pupils.

Murdoch Analytic Sewing Scale for Separate Stitches. This scale was designed to supplement the Murdoch Sewing Scale, and is primarily for the pupils' use. The five stitches measured are running, backstitch, overcasting, combination, and hemming. Full directions are furnished, and norms from the Murdoch Sewing Scale are available for this series.

V. SURVEY TESTS

Illinois Examination. This examination consists of the Illinois General Intelligence Scale, the Monroe General Survey Scale in Arithmetic, and the Monroe Standardized Silent Reading Test, Revised, published as a single test of scholastic ability and achievement. The dual aspect of this battery enables calculation of achievement quotients in arithmetic and reading.

Lippincott-Chapman Classroom Products Survey Test. This test is designed to survey the work of the schools in reading and arithmetic. It consists of two reading tests and two arithmetic tests, covering Grades V to VIII. The arithmetic tests are similar to the Woody-McCall Mixed Fundamentals, and the reading tests to the Monroe Standardized Silent Reading Test, Revised. Table 57 contains data on the two reading tests of the Lippincott-Chapman battery.

Otis Classification Test. This test affords a combined measure of mental ability and educational achievement for Grades IV to VIII.

Stanford Achievement Test. The Arithmetic, Reading, and Dictation Exercises in this test have already been touched upon. The aim of the test as set forth by its authors "was to construct a battery of tests which would cover practically all the curriculum from Grades II to VIII, which would be easy to administer, which would not be too time-consuming, which would yield consistent results (have a high reliability), and which would have the greatest possible validity as a basis for the grading and classification of pupils." It may be said that the test authors, through very extensive experimentation and revision, have come remarkably close to accomplishing these ends.

The nine tests comprising the Advanced Examination are the following:

1. Reading: Paragraph Meaning
2. Reading: Sentence Meaning
3. Reading: Word Meaning
4. Arithmetic: Computation
5. Arithmetic: Reasoning
6. Nature Study and Science
7. History and Literature
8. Language Usage
9. Dictation Exercise

The administration of the test presents no difficulties, and scoring is completely objective. The time required for the complete examination is 2 hours and 20 minutes. It is recommended that the test be given in three sittings, the first covering the Reading tests (48 minutes), the second the Arithmetic tests (42 minutes), and the third Nature Study and Science, History and Literature, Language Usage, and Dictation (50 minutes). Experiments have shown that the performance of mental and educational tasks is not adversely affected during periods of this length.

Extensive norms for the Stanford Achievement Test, including norms for each sub-test, are given in the Manual of Directions. In addition to raw scores, age equivalents are given for each test. This enables computation of the pupil's subject age and general educational age. The reliability of the tests published separately has already been indicated. The reliability of the whole examination for ages 7 to 15 is given in Table 61.

TABLE 61
STANFORD ACHIEVEMENT TEST

AGES	RELIABILITY COEFFICIENTS	P.E. OF EDUCATIONAL AGES IN MONTHS
7	.98	1.1
8	.98	1.6
9	.98	1.6
10	.99	1.5
11	.97	2.0
12	.98	2.1
13	.99	1.9
14	.98	2.1
15	.98	2.1

VI. INTELLIGENCE TESTS

Illinois General Intelligence Scale. This scale was developed from material standardized in the Army Alpha investigations. It is designed for Grades III to VIII. The Illinois Scale has a reliability of .92 for the entire range of talent, and was found to correlate .90 with a composite of 13 intelligence tests.

National Intelligence Tests. The National Intelligence Tests have been among the most widely used of those which were based on the Army Tests. It is comparable to the Illinois Scale in content, reliability, and validity.

Test Materials

MATHEMATICS TESTS

Buckingham Scale for Problems in Arithmetic, Division III. By B. R. BUCKINGHAM. 80 cents per 100. Specimen set, 8 cents. Public School Publishing Company, Bloomington, Illinois.

Compass Diagnostic Tests, Form A. By G. M. RUCH, F. B. KNIGHT, H. A. GREENE, and J. W. STUDEBAKER. The following tests are available:

- Test 1. Addition of Whole Numbers, \$1.00 per 100.
- Test 2. Subtraction of Whole Numbers, \$1.00 per 100.
- Test 3. Multiplication of Whole Numbers, \$2.00 per 100.
- Test 4. Division of Whole Numbers, \$2.00 per 100.
- Test 5. Addition of Fractions and Mixed Numbers, \$2.00 per 100.
- Test 6. Subtraction of Fractions and Mixed Numbers, \$1.00 per 100.
- Test 7. Multiplication of Fractions and Mixed Numbers, \$1.00 per 100.
- Test 8. Division of Fractions and Mixed Numbers, \$2.00 per 100.
- Test 9. Addition, Subtraction, and Multiplication of Decimals, \$2.00 per 100.
- Test 10. Division of Decimals, \$2.00 per 100.
- Test 11. Addition and Subtraction of Denominate Numbers, \$2.00 per 100.
- Test 12. Multiplication and Division of Denominate Numbers, \$2.00 per 100.
- Test 13. Mensuration, \$4.00 per 100.
- Test 14. The Basic Facts of Percentage, \$4.00 per 100.
- Test 15. Interest and Business Forms, \$4.00 per 100.
- Test 16. Definitions, Rules, and Vocabulary of Arithmetic, \$2.00 per 100.
- Test 17. Problem Analysis, Elementary, \$5.00 per 100.
- Test 18. Problem Analysis, Advanced, \$5.00 per 100.
- Test 19. General Problem Scale, Elementary, \$1.00 per 100.
- Test 20. General Problem Scale, Advanced, \$1.00 per 100. Manual of Directions, 20 cents.

Bureau of Educational Research and Service, University of Iowa, Iowa City, Iowa.

Courtis Standard Research Tests, Series B. By S. A. COURTIS. \$1.00 per 100. S. A. Courtis, School of Education, University of Michigan, Ann Arbor, Michigan.

Monroe General Survey Arithmetic Test. By W. S. MONROE. Scale II, Forms 1, 2, and 3, each form \$1.00 per 100. Specimen set, 15 cents. Public School Publishing Company, Bloomington, Illinois.

Monroe Standardized Reasoning Test in Arithmetic, Test III. By W. S. MONROE. 80 cents per 100. Specimen set, 8 cents. Public School Publishing Company, Bloomington, Illinois.

Otis Arithmetic Reasoning Test. By A. S. OTIS. Forms A and B, each 40 cents per package of 25, with Directions, Key, and Class Record. Specimen set, 10 cents. World Book Company, Yonkers-on-Hudson, New York.

Spencer Diagnostic Tests in Arithmetic, III. By P. L. SPENCER. \$2.00 per 100, including Record Sheets and Score Cards. Specimen set, 10 cents. Bureau of Administrative Research, University of Cincinnati, Cincinnati, Ohio.

Stanford Achievement Test: Arithmetic Examination. By T. L. KELLEY, G. M. RUCH, and L. M. TERMAN. Forms A and B, each \$1.00 per package of 25, including Key and Class Record. Manual of Directions (64 pages), 30 cents. World Book Company, Yonkers-on-Hudson, N. Y.

Stone Reasoning Test. By C. W. STONE. Forms I and II, each form 40 cents per 100, \$3.25 per 1000. Manual of Directions, 65 cents. Bureau of Publications, Teachers College, Columbia University, New York.

Woody Arithmetic Scales. By C. WOODY. Four scales of Series A are printed separately. Each scale, 50 cents per 100, \$4.25 per 1000. Series B is printed on a 4-page folder, \$1.50 per 100, \$13.00 per 1000. A Direction Sheet is furnished. Teachers College, Columbia University, New York.

Woody-McCall Mixed Fundamentals. By C. WOODY and W. A. McCALL. Each form, 60 cents per 100, \$5.50 per 1000, including Direction Sheet. Sample set, 20 cents. Teachers College, Columbia University.

Wisconsin Inventory Tests in Arithmetic.

Test I. 100 Combinations in Simple Addition.

Test II. 100 Combinations in Subtraction.

Test III. 100 Combinations in Multiplication.

Test IV. 100 Combinations in Division.

Test V. Higher Decade Addition Facts.

Test VI. Combinations for Carrying in Multiplication.

Test VII. Zero Quotient Combinations in Short Division.

Test VIII. Major Difficulties in Long Division.

Each test, \$1.00 per 100. Specimen set, 10 cents. Public School Publishing Company, Bloomington, Illinois.

ENGLISH

Briggs English Form Test. By T. H. BRIGGS. Forms Alpha and Beta, each form 80 cents per 100, \$7.50 per 1000. Direction Sheet and Scoring stencil, 10 cents. Teachers College, Columbia University.

Charters Diagnostic Language and Grammar Tests. By W. W. CHARTERS. Diagnostic Language Tests, Forms 1 and 2, each 80 cents per 100. Sample set, 10 cents. Diagnostic Grammar Tests, Forms 1 and 2, each \$1.50 per 100. Specimen set, 10 cents. Public School Publishing Company, Bloomington, Illinois.

Franseen Diagnostic Tests in Language. By C. E. FRANSEEN. Each part, \$2.00 per 100. Specimen set, 10 cents. Bureau of Administrative Research, University of Cincinnati, Cincinnati, Ohio.

Ayres Spelling Scale. By L. G. AYRES. 12 cents per copy. Russell Sage Foundation, New York, N. Y.

Buckingham Extension of the Ayres Spelling Scale. 12 cents per copy, for 3 or more copies. Public School Publishing Company, Bloomington, Illinois.

Iowa Dictation Exercise and Spelling Tests. Form C, 10 cents per copy. Iowa State Spelling List (based on the Iowa Spelling Scale), 2 cents per copy. Bureau of Educational Research and Service, University of Iowa, Iowa City, Iowa.

Chapman-Cook Speed of Reading Test. By J. C. CHAPMAN and S. COOK. \$1.00 per 100. J. B. Lippincott Company, Philadelphia, Pennsylvania.

Gray Standardized Oral Reading Check Test. By W. S. GRAY. Set IV. \$1.50 for all test material needed for 20 pupils. Public School Publishing Company, Bloomington, Illinois.

Monroe Standardized Silent Reading Tests, Revised. By W. S. MONROE. Test II, Forms 1 and 2, each \$1.00 per 100. Specimen set, 15 cents. Public School Publishing Company, Bloomington, Illinois.

Stanford Achievement Test: Reading Examination. By T. L. KELLEY, G. M. RUCH, and L. M. TERMAN. Forms A and B, each \$1.10 per package of 25, including Key and Class Record. Manual of Directions (64 pages), 30 cents. World Book Company, Yonkers-on-Hudson, New York.

Stone Narrative Reading Tests. By C. R. STONE. \$4.00 per 100. Individual Record Sheet, 60 cents per 100. Public School Publishing Company, Bloomington, Illinois.

Thorndike Test of Word Knowledge. By E. L. THORNDIKE. Each form, \$1.50 per 100, including Manual of Directions. Teachers College, Columbia University, New York.

Willing Scale for Measuring Written Composition. By M. H. WILLING. 6 cents each for 3 or more copies. One scale is needed for each judge. Public School Publishing Company, Bloomington, Illinois.

GEOGRAPHY

Buckingham-Stevenson Place Geography Tests and Buckingham-Stevenson Information-Problem Tests in Geography. By B. R. BUCKINGHAM and P. R. STEVENSON. The Place Geography Tests are dictated by the teacher; 3 forms of each test, 20 cents. One copy of the Information Test is needed for each student; \$2.00 per 100. Specimen set, 15 cents. Public School Publishing Company, Bloomington, Illinois.

Courtis Standard Supervisory Test in Geography. By S. A. COURTIS. Tests A and B, with Forms A and B for each. \$1.00 per 100 for each form.

Bureau of Educational Research and Service, University of Iowa, Iowa City, Iowa.

Gregory-Spencer Geography Tests. By C. A. GREGORY and P. L. SPENCER. Forms A, B, and C, each \$1.00 per package of 25. Specimen set, 10 cents. Bureau of Administrative Research, University of Cincinnati, Cincinnati, Ohio.

Hahn-Lackey Geography Scale. By H. H. HAHN and E. E. LACKEY. One copy needed for each teacher. 15 cents each for 3 or more copies. Public School Publishing Company, Bloomington, Illinois.

Posey-Van Wagenen Geography Scales. By C. J. POSEY and M. J. VAN WAGENEN. Each scale \$1.50 per 100. Teacher's Handbook, 20 cents. Specimen set, 30 cents. Public School Publishing Company, Bloomington, Illinois.

VOCATIONAL SUBJECTS

Ayres Measuring Scale for Handwriting, Gettysburg Edition. By L. G. AYRES. 10 cents per copy. Russell Sage Foundation, New York, N. Y. *Courtis Standard Practice Tests in Handwriting.* By S. A. COURTIS and LENA A. SHAW. Student's Daily Lesson Book, 10 cents; Teacher's Manual, 20 cents; Class Record, 5 cents. Specimen set, 45 cents. World Book Company, Yonkers-on-Hudson, New York.

Freeman Chart for Diagnosing Faults in Handwriting. By F. N. FREEMAN. 30 cents per copy. Houghton Mifflin Company, Boston.

Thorndike Handwriting Scale. For General Merit of Children's Handwriting. By E. L. THORNDIKE. One copy of the scale needed for each teacher. 10 cents per copy in quantities. Bureau of Publications, Teachers College, Columbia University, New York.

Gates-Strang Health Knowledge Test. By A. I. GATES and R. STRANG. Bureau of Publications, Teachers College, Columbia University, New York.

Short Scales for Measuring Habits of Good Citizenship. By L. M. CHASSELL, S. M. UPTON, and C. F. CHASSELL. Scales C, D, G, and H are available. Each scale 50 cents per 100. Bureau of Publications, Teachers College, Columbia University, New York.

Home Economics Information Tests. Prepared by the Household Arts Department of Teachers College. 15 cents per set of 3 booklets. Specimen set, 35 cents. Bureau of Publications, Teachers College, Columbia University, New York.

King-Clark Foods Test. By F. B. KING and H. F. CLARK. Form A, \$1.00 per package of 25, including Manual of Directions, Key, and Class Record. Specimen set, 10 cents. World Book Company, Yonkers-on-Hudson, New York.

Murdoch Scale for Measuring Certain Elements in Hand Sewing. By K. MURDOCH. For each examiner, one copy of the scale, including Manual of Directions, \$1.50.

Murdoch Analytic Sewing Scale for Separate Stitches. By K. MURDOCH.

For the examiner, one copy of the Manual of Directions, 10 cents; for the pupil, one copy of the scale, 25 cents. Bureau of Publications, Teachers College, Columbia University, New York.

SURVEY TESTS

Illinois Examination. By W. S. MONROE and B. R. BUCKINGHAM. Examination II, Forms 1 and 2, each form \$1.00 per 100. Teacher's Handbook, 15 cents. Specimen set, 25 cents. Public School Publishing Company, Bloomington, Illinois.

Lippincott-Chapman Classroom Products Survey Test. By J. C. CHAPMAN. \$3.50 per 100. J. B. Lippincott Company, Philadelphia.

Otis Classification Test. By A. S. OTIS. Forms A and B, each \$1.10 per package of 25. Manual of Directions, 25 cents. Specimen set, 40 cents. World Book Company, Yonkers-on-Hudson, New York.

Stanford Achievement Test: Advanced Examination. By T. L. KELLEY, G. M. RITCH, and L. M. TERMAN. Forms A and B, each form \$1.90 per package of 25, including Key and Class Record. Manual of Directions (64 pages), 30 cents. Specimen set, 60 cents. World Book Company, Yonkers-on-Hudson, New York.

INTELLIGENCE TESTS

Illinois General Intelligence Scale. By W. S. MONROE and B. R. BUCKINGHAM. Forms 1 and 2, each \$2.00 per 100. Sample set, 20 cents. Public School Publishing Company, Bloomington, Illinois.

National Intelligence Tests. By a committee (HAGGERTY, TERMAN, THORNDIKE, WHIPPLE, and YERKES). Scales A and B, 3 forms for each scale, each form \$1.25 per package of 25, including Key and Record Sheet. Manual of Directions, 20 cents. Specimen set, 50 cents. World Book Company, Yonkers-on-Hudson, New York.

Terمان Group Test of Mental Ability. By L. M. TERMAN. Forms A and B, each \$1.20 per package of 25, with Key, Record Sheet, and Manual of Directions. Specimen set, 15 cents. World Book Company, Yonkers-on-Hudson, New York. (Usable in Grades VII and VIII, as well as in the high school.)

References

BRIGGS, T. H. *The Junior High School*. Houghton Mifflin Company, Boston; 1920. 350 pages.

COURTIS, S. A. "The Influence of Certain Social Factors upon Scores in the Stanford Achievement Tests." *Journal of Educational Research*, Vol. XIII, No. 5 (May, 1926), pages 311-324.

CURRENT, W. F., and RITCH, G. M. "Further Studies on the Reliability of Reading Tests." *Journal of Educational Psychology*, September, 1926, pages 476-481.

DAVIS, C. O. *Junior High School Education*. World Book Company, Yonkers-on-Hudson, New York; 1924. 451 pages.

EYESTONE, A. B., and RITCH, G. M. *Studies on Standard Tests in Geography*. To be published.

GATES, A. I. "An Experimental and Statistical Study of Reading and Reading Tests." *Journal of Educational Psychology*, Vol. XII (1921), pages 303-314, 378-391, 445-464.

GLASS, J. M. *Curriculum Practices in the Junior High School and Grades 5 and 6*. Supplementary Educational Monograph, No. 25. University of Chicago, Chicago; November, 1924. 181 pages.

GRAY, W. S. "Summary of Reading Investigations (July 1, 1924 — June 30, 1925)." *Elementary School Journal*, Vol. XXVI, No. 6 (February, 1926), pages 449-459; No. 7 (March, 1926), pages 507-518.

HINES, H. C. *Junior High School Curricula*. The Macmillan Company, New York; 1924. 188 pages.

MURDOCH, K. "The Measurement of Certain Elements of Hand Sewing." *Teachers College Contributions to Education*, No. 103 (1919). Columbia University, New York.

New York Survey of Junior High Schools. Board of Education, New York; 1924. 257 pages.

OTIS, ARTHUR S. "The Making of a Classification Test." *Contributions to Education*, Vol. I, Chapter XIV. World Book Company, Yonkers-on-Hudson, New York; 1924.

PECHSTEIN, L. A., and McGREGOR, A. L. *Psychology of the Junior High School Pupil*. Houghton Mifflin Company, Boston; 1924. 280 pages.

Report of the Committee of Fifteen on Secondary Education in California. California High School Teachers' Association; 1924. 405 pages.

Survey of the Junior High Schools of New York (City). Board of Education, New York; 1924. 257 pages.

WILLING, M. H. "Individual Diagnosis in Written Composition." *Journal of Educational Research*, Vol. XIII, No. 2 (February, 1926), pages 77-89.

WITHAM, E. C. "Standard Geography Tests." *American School Board Journal*, Vol. LXXI (November, 1925), pages 51, 52, 129.

PART THREE
INFORMAL OBJECTIVE EXAMINATION
METHODS

CHAPTER FOURTEEN

THE RÔLE OF INFORMAL OBJECTIVE EXAMINATIONS IN HIGH SCHOOL INSTRUCTION

Introduction. In the complete program of instruction and measurement in the high school, there are many situations where standard tests are not practicable. Due to the cost, time requirements, non-adaptability, etc., of validated and standardized tests, less formal examination methods are highly desirable for daily, monthly, or even semester tests and examinations. The traditional examination has always filled this need until recent times, when the reliability of the usual discussion or essay type of examination has been called into serious question. In fact, a considerable amount of experimental evidence has been brought forward which tends to discredit the usual examination for many of its avowed purposes. We shall have occasion to examine some of this evidence in the three following chapters.

The relation between standardized and unstandardized tests. The increasing use of such examination techniques as the true-false, completion, multiple-response, and matching tests has led some persons to hold that there is a distinct conflict of purpose between such informal test methods and the standard educational test. This is far from the truth, as time is certain to demonstrate. Each of these methods has its legitimate rôle in the measurement of instruction, and the two methods are to be regarded as supplementary and not antagonistic.

The real rivalry will take place between the newer objective examination and the traditional essay or discussion test. Even then, we shall probably always have three more or less distinct types of measures in use in our schools; viz.,

- (1) The traditional discussion examination;
- (2) The informal objective test;
- (3) The standard educational test (and the standardized mental examination).

Although Chapters XIV, XV, and XVI will concern themselves principally with informal objective examinations more or less to the exclusion of the usual written examination, it is nevertheless true that the traditional type of examination will probably always be with us as a regular classroom procedure. However, it is quite likely that its use will be confined to those school subjects and those aspects of all school subjects which cannot be measured in purely objective fashion. For examinations of factual character, it seems a fair prediction to assert that the traditional examination will gradually give way to the informal objective examination and the standard test.

Limitations of the traditional examination. The three principal functions of examinations center about (a) measurement, (b) motivation, and (c) training in the use of the English language. The order of stating these purposes is thought to be the order of importance by the authors. These three aims of tests and examinations will, however, be discussed in a different sequence.

It is quite obvious that neither the standard test nor the informal objective examination lays any claim to the function of language training. In an instructional sense this is an avowed limitation; from the standpoint of pure measurement it is a source of strength, since the evaluation of school work from written examinations is a highly subjective process fraught with error.

When 91 teachers grade the same pupil's answer to a question in geography and show a range of marks of from 2 to 20 out of a possible 20 points, it is a fair question to ask

whether either linguistic training or factual mastery has been measured. Yet such a result has actually been found experimentally.¹

The usefulness of examinations of the usual sort in developing such abilities as organization of thought, reasoning, and good literary expression cannot be held to be totally negligible, but it is a safe assertion that more or less radical changes will have to be made in examination practices to attain these desired ends. Some of the necessary reforms are the following:

(1) The time allowed per question for answering the examination must be materially increased — at least doubled or trebled. The “hustle and bustle” of present written examinations alone will prevent the realization of the ends of language training through the examination. Worse still, not a few careful thinkers are convinced that the usual 10-question-per-hour examination actually operates to stimulate the worst efforts of the pupil from a linguistic point of view.

(2) Examinations of two or three topics per hour must be set for the pupils with the full realization on the part of both teacher and pupil that *form, not content*, will be the principal if not the sole basis upon which the papers will be graded. The same standard should prevail as in theme-writing in the regular English classroom.

(3) The measurement of factual mastery must be divorced more or less completely from such an examination and handled by a more objective and reliable technique of testing. It will be for the future to decide whether such techniques will be the true-false, the completion, the multiple-choice, or yet other methods which time will bring forth. At any rate, there is plenty of trustworthy evidence which shows that the traditional examination fails to be a reliable means of measurement of facts and formal skills.

¹ Ruch, G. M., *The Improvement of the Written Examination* (Scott Foresman & Co., 1924), pages 56-62.

As far as the three types of examinations which have been developed to date are concerned, there is no experimental evidence which will evaluate their respective merits in the motivation of learning. Any conclusions which might be drawn on this issue must rest on *a priori* grounds. It would appear to be reasonable to suppose that the motivating effects of tests and examinations would increase directly with the definiteness and searchingness of the final examination. In this case the objective examination and the standard test would very likely prove somewhat superior to the ordinary examination, open as it is to bluffing, evasions, and the mere writing of words.

It is on the ground of measurement that the traditional examination has been most often attacked. Due to its subjectivity, it has been shown to be highly unreliable for at least two principal reasons; viz.,

- (1) The subjectivity of the scoring, and
- (2) The limited sampling permitted by a 5- or 10-question examination.

These two defects of the traditional examination will be considered at some length.

The subjectivity of the traditional examination. Starch and Elliott,¹ as early as 1912, carried out several investigations which established the fact that the grading of examinations is a highly subjective process. A pupil's paper in geometry was duplicated and sent to 115 experienced teachers for grading. When the results were tabulated, it was found that a range of from 28 per cent to 92 per cent existed in the marks assigned by this group of teachers. A paper in English, graded by 142 teachers, yielded a range in marks of from 50 per cent to 98 per cent.

¹ *School Review*, Vol. XX, pages 442-457; Vol. XXI, pages 254-259; Vol. XXVI, pages 676-681.

Such disagreements show clearly that neither of the two pupils writing these papers could be said to be measured. In fact, the difference between the dunce and the class star would seem to lie, not in the knowledge which he possesses, but rather in the teacher who happened to be in charge of his class!

One of the authors has repeated this work of Starch and Elliott and has confirmed its accuracy in all important respects.¹

The table and graph which follow are taken from the source just cited (pages 59-60) and show the variation of subjective opinion in the marking of three papers in United States history — the papers being selected as the best, median, and poorest papers, respectively, in a class of about 30 pupils, according to the marks of the original or regular teacher.

Table 62 shows not only a systematic error in the direction of leniency on the part of the original grader of these three papers, but also a more marked lack of agreement among the 115 teachers who regraded the papers. The following conclusions may be drawn from inspection of Table 62 and Figure 1:

- (1) The original marks were systematically too high by from 10 to 18 points if the average marks of 115 teachers are to be trusted as approximating the truth of the matter.
- (2) Variations among the 115 teachers are at times great enough to cause a pupil to be rated as "excellent" by one teacher and as "failed" by another.
- (3) Certain teachers marked the poorest paper in the class higher than certain other teachers rated the best paper, and vice versa.

¹ Ruch, G. M., *The Improvement of the Written Examination* (Scott, Foresman & Co., 1924), pages 51-62.

TABLE 62

ORIGINAL MARKS AND MARKS ASSIGNED BY 115 TEACHERS WHO
REGRADED THREE PAPERS FROM A CLASS IN AMERICAN HISTORY

MARK	PAPER 1	PAPER 2	PAPER 3
100	6	—	—
95	33	—	—
90	32	1	—
85	22	12	1
80	15	13	4
75	6	29	9
70	1	18	5
65	—	16	20
60	—	12	16
55	—	9	19
50	—	3	13
45	—	2	8
40	—	—	14
35	—	—	5
30	—	—	0
25	—	—	1
Number . . .	115	115	115
Mean . . .	88.7	70.3	56.6
Original mark .	100	88	67
Standard Devia- tion . . .	6.6	9.9	12.3

- (4) The expedient of having 115 teachers grade each paper failed to establish the true worth of the paper unless we are willing to rest the case on statistical grounds — i.e., accept the average as the truth.
- (5) The approximate worth of such papers could only be obtained by having at least 10 or more teachers mark the same papers, and this is obviously not a practical solution of the problem.

There are many other interesting studies on the fallibility of teachers' marks which can be consulted by the reader.

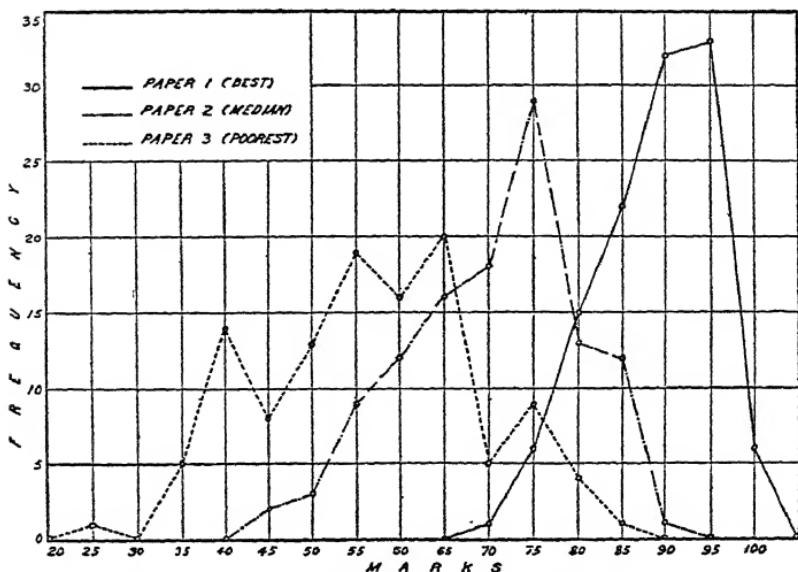


FIG. 1. Curves showing the distributions of the marks of 115 teachers who regraded the three history papers in Table 62.

through the use of the references at the end of Part III. The books and articles of Starch and Elliott, Monroe, Wood, Toops, Paterson, Russell, and Ruch contain a great deal of evidence on points passed over very superficially in the preceding pages.

The second weakness of the traditional examination will next be considered; viz., unreliability due to limited sampling.

Unreliability of the traditional examination due to limited sampling. Another major weakness of the usual examination centers about the fact that it is usually but a 5- or 10-question sample of the course or subject under consideration. This is in sharp contrast with the typical standard test or objective examination, which is made up of 50, 100, 150, or more questions. Obviously the questions are narrower and less inclusive in the case of the two objective types of test. The use of 5 or 10 broad discussion questions can be termed

an *intensive sampling* in contrast with the *extensive sampling* provided by the informal objective examination or the standard test. The traditional examination explores a few topics thoroughly, and the two objective types of tests sample many topics less completely.

Any test or examination must be thought of as a sampling of abilities rather than complete measurement. It is possible that, in a very narrow ability like the 100 basic multiplication facts, an examination might approach *total measurement*; but on the whole the truth of such a statement is to be doubted, since even the presentation of these 100 multiplication facts time after time in varying order will yield slightly different per cents of correct responses by a pupil, if for no other reasons than haste, carelessness, lapses of attention, misreading, accidental miswriting, etc.

There are certain implications resident in the method of drawing up 5 or 10 final examination questions, and these are seldom borne consciously in mind by the teacher. A few brief statements make clear some of the assumptions underlying examination construction.

- (1) Any examination is a limited sampling.
- (2) The actual questions selected depend in part upon the teacher's point of view at the moment. On a different day, a somewhat different lot of questions might be drafted.
- (3) The actual lot of questions drawn up will ordinarily yield a different average mark, a different range of marks, and different individual marks from those of a second set of questions by the same teacher at a different time, or by a different teacher handling the same course.
- (4) If a teacher made up two sets of 10 final examination questions and gave both examinations to the same class,

the results, pupil by pupil, will not agree in their numerical statements or even in the rank orders of the pupils on the two examinations.

- (5) Other things being equal, the longer the examination (in terms of numbers of questions or of actual working time), the more valid and reliable the results.
- (6) The statement of a pupil's mark as 85 per cent can at best mean merely that he answered 85 per cent of the questions asked; never that he knows 85 per cent of the subject, since he probably was not examined over more than 5 per cent to 20 per cent of the subject matter.
- (7) The possible errors in ascertaining the pupil's accomplishment by means of an examination will be further complicated by the personal equation of the reader of the paper.
- (8) With a few-question examination, errors due to absences of the pupil from school, misunderstanding of the phrasing of a question, etc., will tend to be larger than in the many-question examination with its wider sampling.

The foregoing are but a few of many considerations which should be conscious in our thinking about examination practices. Some of these overlap the matter of subjectivity as well. Perhaps the best way to illustrate the full import of both subjectivity and sampling as factors in written examinations is to describe at some length an actual investigation which covers both issues.

The following pages present a brief summary of some findings from an investigation of the state eighth-grade examinations as set by the state departments of public instruction in many states of the Union.¹

¹ Ruch, G. M., et al. *Objective Examination Methods in the Social Studies* (Scott, Foresman & Co., 1926; 116 pages), especially pages 6-22.

Table 63 shows the result of administering 16 eighth-grade state examinations from 10 different states to a number of groups of pupils, each pupil writing on both the 1923 and the 1924 sets of questions. Also, two experienced teachers of the particular subjects read each paper, thus giving four marks on each of the 952 pupils, as follows:

- (1) The 1923 set read by Teacher No. 1
- (2) The 1923 set read by Teacher No. 2
- (3) The 1924 set read by Teacher No. 1
- (4) The 1924 set read by Teacher No. 2

Since there were four separate marks for each of the pupils, it was possible to compare the amount of agreement among the four sets of marks by correlation methods, six such correlation coefficients being possible for each of the 16 examinations. The six columns of Table 63 group themselves into three general situations to be described later. Notice that averages are given for the three such pairs of columns as well as for the separate columns.

The names of the states have been replaced by key numbers.

The following information is necessary in reading the table:

- Column (1) Correlation of marks of Teachers 1 and 2 on the 1923 examination
- Column (2) Correlation of marks of Teachers 1 and 2 on the 1924 examination
- Column (3) Correlation of Teacher 1's marks on 1923 and 1924 examinations
- Column (4) Correlation of Teacher 2's marks on 1923 and 1924 examinations
- Column (5) Correlation of marks given by Teacher 1 on the 1923 examination with the marks of Teacher 2 on the 1924 examination

TABLE 63

RELIABILITY COEFFICIENTS OF 16 EIGHTH-GRADE STATE EXAMINATIONS,
YEAR (1923) AGAINST YEAR (1924), AND TEACHER 1 AGAINST TEACHER 2

No.	KEY	SUBJECT	(1)	(2)	(3)	(4)	(5)	(6)	Pop.
1	G-2	Ele. Citizenship .	.45	.21	-.05	.46	.34	-.26	102
2	I-1	U. S. History .	.60	.43	.16	.41	.23	.17	31
3	J-1	U. S. History .	.47	.30	.44	.73	.25	.22	32
4	F-3	Geography . .	.58	.39	.37	.22	.17	.55	36
5	F-1	U. S. History .	.89	.99	.67	.64	.67	.69	94
6	D-2	Civics82	.88	.22	.25	.33	.23	36
7	I-3	Geography40	.88	.32	.48	.29	.41	32
8	M-2	Civics80	.82	.47	.55	.46	.52	61
9	D-1	U. S. History .	.81	.22	.54	.65	.49	.35	34
10	L-1	U. S. History .	.79	.57	.73	.48	.45	.66	107
11	A-1	U. S. History .	.81	.85	.66	.71	.56	.65	42
12	B-1	U. S. History .	.53	.58	.36	.34	.27	.41	97
13	B-2	Civics63	.20	.36	-.18	-.06	.25	82
14	K-1	U. S. History .	.93	.91	.56	.67	.68	.51	99
15	I-2	Civics81	.53	.37	.27	.52	.46	35
16	E-1	U. S. History .	.75	.12	.26	.59	.60	.19	32
Averages69	.56		.40	.45		
Averages by pairs of columns			.62			.43		.38	(952)

Column (6) Correlation of marks given by Teacher 1 on the 1924 examination with the marks of Teacher 2 on the 1923 examination

As already mentioned, the six columns of Table 63 group themselves in three general situations whose average correlations are .62, .43, and .38, respectively.

Situation I (Columns 1 and 2)

Here the examination is the constant and the marker of the papers is the variable. This situation allows the sub-

jectivity of scoring to be the main variable, since two independent readers marked the *same papers*. There is no statistical reason, apart from subjective differences, why all of the correlations in Columns (1) and (2) could not equal 1.00. Actually, they average .62. A correlation of .62 means that the agreement of the two independent readings, on the average, was about 22 per cent better than that accountable for by chance (zero correlation). We can describe this situation in actual practice as being typical of what might happen if the regularly appointed reader could not serve and a substitute had to be secured. By and large, in correlational terms, agreements of about .60 to .65 would seem to be the average expectancy.

Situation II (Columns 3 and 4)

Here the reader of the papers is the constant and the examination is the variable. The factor of subjectivity still enters, but in its minimum effect, since the *same reader* scored two different examinations. Such correlations must theoretically always be less than unity (1.00) because the limited sampling of 5- or 10-question examinations results in unreliability, and also because there remains a certain amount of subjectivity even when the same person reads both sets of papers. It seems reasonable to expect that the correlations in these two columns will average smaller than for the two preceding columns; and such is the case, the average here being .43, an agreement which represents about 10 per cent improvement over chance assignment of marks. Probably the main reason for lack of higher correlation in these marks from the 1923 and 1924 examinations is to be found in the unreliability of small samples. It is obvious that it makes considerable difference in which year pupils complete their elementary education and present themselves for examination. To place this situation in a more practical

setting, we can suppose that a given state keeps the same reader year after year. A pupil contemplating taking the 1923 examination is prevented by illness and must wait over until the next year. On the average, the effect of such a situation was represented by the average correlation of .43.

Situation III (Columns 5 and 6)

Here both the examination and the scorer are variables. This third situation is the one which parallels actual conditions in most states (giving uniform examinations) most closely; i.e., the questions differ from year to year, and a different reader is appointed each year. A pupil becoming ill in May, 1923, and forced to wait until May, 1924, would have his paper marked by Miss Smith instead of Mr. Brown, who was the official reader for 1923. The average agreement under Situation III was represented by a correlation of .38, roughly 8 per cent better than chance assignment of marks.¹

An even simpler treatment of the results of the same investigation can be carried out by comparing the differences in the average marks on the 16 examinations. As before, since there are 4 marks on each pupil, there are 6 possible sets of differences which can be stated. Table 64, which follows, presents the means, and the largest, smallest, and average differences in the average marks assigned to each of the 16 sets of papers. The 4 marks on each paper can be designated by the letters A, B, C, and D, as follows:

¹ Throughout the foregoing discussion "chance assignment of marks" is defined as some such practice as the placing of the actual marks from the examination on bits of paper and then shaking them up in a hat, the pupils receiving their marks by drawing from the hat at random, just as in drawing lots in a game. The statistical basis of the statement of "per cents better than chance assignment" is to be found in the formula for the standard error of estimate; viz., $\sigma_{1.2} = \sigma_1 \sqrt{1 - r^2}$, or, more simply, *Lack of agreement varies as $\sqrt{1 - r^2}$* .

TABLE 64

THE AVERAGE MARKS ASSIGNED BY TWO DIFFERENT TEACHERS TO
BOTH THE 1923 AND THE 1924 EXAMINATIONS

No.	KEY No.	A	B	C	D
		1923 Exam. Scorer 1	1923 Exam. Scorer 2	1924 Exam. Scorer 1	1924 Exam. Scorer 2
1	G-2	67.5	82.0	73.0	70.4
2	I-1	71.7	67.1	57.0	71.0
3	J-1	68.1	43.6	64.9	45.6
4	F-3	70.0	54.8	70.3	69.9
5	F-1	47.7	45.3	51.4	41.9
6	D-2	55.7	47.5	65.6	59.1
7	I-3	51.0	62.3	50.4	68.9
8	M-2	51.0	48.6	48.5	42.9
9	D-1	38.3	34.4	48.5	30.7
10	L-1	49.3	56.5	38.3	65.3
11	A-1	42.5	28.1	25.3	18.6
12	B-1	14.4	7.7	24.4	25.3
13	B-2	29.3	26.0	24.8	11.5
14	K-1	48.1	59.0	61.3	64.7
15	I-2	38.3	41.4	68.1	58.0
16	E-1	21.0	26.9	8.6	12.4

SUMMARY OF DIFFERENCES

	(A-B)	(C-D)	(A-C)	(B-D)	(A-D)	(A-C)
Average Difference	. 8.6	9.4	9.2	9.9	11.4	12.0
Largest Difference	. 24.5	17.6	29.8	27.0	23.9	26.8
Smallest Difference	. 2.4	2.0	0.3	0.4	0.1	0.2

A Average mark for 1923 examination given by Teacher No. 1

B Average mark for 1923 examination given by Teacher No. 2

- C Average mark for 1924 examination given by Teacher No. 1
- D Average mark for 1924 examination given by Teacher No. 2

The summary of differences at the bottom of Table 64 is self-explanatory, but it should be remembered that we are here dealing with differences in *averages*, not differences in the marks of *individual pupils*. The variations of the marks of individual pupils from the 1923 examination to the 1924, and from teacher to teacher, are obviously often much greater (or much smaller) than the differences in the average marks of whole groups of pupils. Not far from half of the variations in individual marks will, of course, be as large as or larger than the values shown in the "Summary of Differences."

Many other studies might be reviewed which show the limitations of the traditional examination, but it will be more profitable to turn to the consideration of constructive suggestions for remedying conditions. Chapter XV presents samples of the various types of informal objective examinations which are at present coming into widespread use.

CHAPTER FIFTEEN

TYPES AND CHARACTERISTICS OF OBJECTIVE EXAMINATIONS

Classification. A simple classification of objective tests can be made as follows:

I. Recall Types:

- (1) Simple recall questions
- (2) Completion exercises

II. Recognition Types:

- (1) Multiple response
- (2) True-False
- (3) Matching exercises
- (4) Best answer or judgment tests
- (5) Identification exercises
- (6) Rearrangement tests

This list does not include all of the known types of objective tests such as would be found by compilations of educational and mental test methods. There are also numerous variates of each of these type forms.

The pages which immediately follow give illustrative materials for all of the types of objective examination in common use. In a later section these various types are described as to their special characteristics, special advantages, and major limitations. Many of the samples are taken from well-known educational tests, a fact which emphasizes the common parentage of these two techniques of educational measurement.

Advantages and limitations of the various forms of objective examination techniques. There are a number of characterizations which can be made with some degree of assurance which will guide the teacher in the choice of the type of test to be employed for particular purposes.

RECALL OR COMPLETION TESTS

I. *Advantages:*

- (1) Relative freedom from the effects of guessing or chance (probably equal or superior to the traditional examination on this score)
- (2) Naturalness of mechanical form (closely parallel to the more or less typical question, "Boyle's law deals with changes in gas — what?")
- (3) Allows some freedom of expression and thought (in fact, Ebbinghaus originated the type as a measure of intelligence)
- (4) Better adapted, perhaps, to thought questions than many other forms

II. *Limitations:*

- (1) Not purely objective
- (2) Takes more time than many other forms
- (3) The use of too many blanks within the same item tends to reduce it to a puzzle situation or intelligence test
- (4) Somewhat slower and more difficult to score than many other objective types

MULTIPLE-RESPONSE TESTS

I. *Advantages:*

- (1) Purely objective in scoring
- (2) Very rapidly scored, especially when the responses are numbered and the answering is done by numbers (this is the most rapid scoring technique yet developed)
- (3) Can be used as a judgment test, especially when the alternative responses are somewhat lengthy statements and present varying degrees of truth or plausibility

- (4) Is easier to prepare than the completion type
- (5) Guessing effects can be minimized by using from four to seven responses

II. *Limitations:*

- (1) Tends to be highly factual
- (2) Dangers of more than one correct or nearly correct answer
- (3) More or less open to marked guessing unless 5, 7, or more alternative answers are presented
- (4) Often difficult to secure more than two or three plausible answers (although, to the uninformed pupil, almost any answer might appear plausible)

TRUE-FALSE TESTS**I. *Advantages:***

- (1) Wide applicability
- (2) Perfectly objective in scoring
- (3) Fairly easy to construct (although the difficulties of securing defensible statements are greater than is commonly supposed)
- (4) Great rapidity in which items can be answered
- (5) Extensiveness of sampling possible in limited time

II. *Limitations:*

- (1) Much important subject matter is neither absolutely true nor false
- (2) Open to marked chance effects
- (3) Scores probably need correction for chance
- (4) Dangers of ambiguity of statement

MATCHING EXERCISES**I. *Advantages:***

- (1) Completely objective
- (2) Can be used as either judgment or factual tests

II. *Limitations.*

- (1) If pairs to be matched are few in number, guessing enters
- (2) If sufficient numbers of pairs are used per block or unit to minimize chance effects, there is some loss of time in the matching process
- (3) In chronological situations, such as are common in history teaching, large blocks or units are likely to throw dates so close together that the degree of discrimination required cannot be defended upon grounds of social utility

BEST ANSWER OR JUDGMENT TESTS

I. *Advantages:*

- (1) Has all the general advantages of the multiple-response test, of which it is to be regarded as a variate
- (2) The vertical arrangement of the responses permits the use of long response statements, thus allowing wide use of the method for thought and judgment questions

II. *Limitations:*

- (1) Requires much space for mimeographing
- (2) Because of space requirements, if used for reasoning purposes, the number of responses must be kept to three to five, thus allowing chance to enter

IDENTIFICATION EXERCISES

I. *Advantages:*

- (1) Allows the testing of ability to apply principles to concrete situations or identify examples of such applications of principles

II. *Limitations:*

- (1) Costly in space
- (2) Limited to certain subjects in which principles and generalizations have many specific applications

REARRANGEMENT TESTS**I. *Advantages:***

- (1) Chiefly useful in subjects involving chronological relations; e.g., history

II. *Limitations:*

- (1) Difficult to score at times, since numerous displacements of the correct order of arrangement are possible, and in varying degrees of error

Advantages and limitations of objective tests in general. In addition to the specific characterizations just given, there are a number of advantages and weaknesses in all objective test methods. A few of these follow.

ADVANTAGES OF OBJECTIVE EXAMINATIONS

1. Objective tests can be made quite or nearly perfectly objective, thus eliminating most of the unreliability due to personal opinion.
2. Objective tests permit a much more extensive sampling per unit of time than do traditional examinations.
3. Objective tests force the child to react to the facts and ideas which the teacher deems important rather than allow the pupil to choose the line of battle. The teacher who has noted the facility with which the pupil can distort the ordinary question to make it fit the knowledge he happens to have retained will recognize the advantages of the more mechanical form of the objective test.
4. The scoring is quickly and accurately carried out, by clerical help if need be.

5. The time ordinarily devoted to the correction of papers can, in large part, be devoted to the preparation of the examination. With the traditional examination, it often happens that the questions can be made out in 10 to 15 minutes but it requires 2 to 3 hours to read the papers. With the objective examination, the ratio is reversed. The objective examination will require, on the whole, more time to construct, but the added time is devoted to the more interesting and fundamental phase of examining — viz., the actual construction of the test materials.

6. Experimentation has shown that well-constructed objective examinations are more valid and reliable than the essay type of test per unit of actual working time.

7. Informal objective examinations are more adaptable for small units of subject matter and to local conditions than are standard tests.

8. Informal objective tests can be used not only for testing purposes but for instructional, diagnostic, and remedial exercises.

LIMITATIONS OF OBJECTIVE EXAMINATIONS

1. Objective examinations provide little or no opportunity for training in organization and expression of thought.

2. There is always a tendency for objective tests to become highly factual and place a premium on memory alone.

3. Guessing effects disturb the accuracy of results to greater or less degree in all such tests unless many alternative responses are offered.

4. Objective tests require the use of the mimeograph if entirely satisfactory results are to be obtained. True-false and certain other forms of statements may be dictated, but this practice has decided limitations.

Illustrations of the more common types of objective examinations. The remaining pages of this chapter will be

devoted to the presentation of a few short samples of objective tests.

SIMPLE RECALL TESTS

DIRECTIONS. Write a word or short phrase on each blank line which will make each statement true.

- 1 Eli Whitney is noted for his invention of the
.....
- 2 The first permanent English settlement in America
was at
.....
- 3 Every President of the United States upon coming
into office now chooses ten Secretaries to form his
.....
- 4 The fundamental economic cause of the Civil War was
.....
- 5 In 1820 Maine was admitted as a free state and Mis-
souri as a slave state under the provisions of the
.....

* * * * *

- 46 President Wilson said, "The world must be made
safe for
.....
- 47 The United States Court ruled that no law applying to
commerce carried on between two or more states could
be passed by any
.....
- 48 The Acts of 1867 by which Congress forced negro
suffrage upon the South were called the
.....
- 49 Large corporations, when consolidated, are popularly
called
.....
- 50 The war which has most generally been condemned by
American historians was our war with
.....

In connection with the foregoing test, note the vertical alignment of all answers. This device facilitates the scoring very greatly, since a strip bearing the acceptable responses may be placed directly at the left of the column of pupils' answers for comparison.

COMPLETION TESTS¹

DIRECTIONS. In each of the paragraphs below write in the words that have been left out. Try to find the word for each blank that makes the best sense.

The term "nitrogen cycle" has often been used to describe the series of transformations which the chemical element nitrogen undergoes in its relations to living matter. are completely lacking in ability to use the free nitrogen of the, although there is an almost unlimited supply of it. The same is true of most, although certain families like the possess structures on their known as which contain the so-called bacteria. These bacteria "fix" the free nitrogen; i.e., they change it into the form of, which, being soluble, are absorbed by the roots of plants and used in the production of Animals in turn eat the plants and further elaborate these compounds. Upon the death of the animals the bacteria of break up these compounds into simpler ones, like ammonia, and finally into The last types of bacteria are called bacteria.

MULTIPLE-RESPONSE TESTS

I. A 4-Response Test from an Examination in Woodwork

DIRECTIONS. Underline the word that makes the correct answer.

- 1 Varnish is thinned with gasoline linseed oil turpentine alcohol
- 2 A plane about nine inches long is called a smoothing plane jack plane fore plane jointer plane
- 3 To bore a $\frac{1}{4}$ -inch hole with a gimlet bit, you would use a bit with the number 10 12 8 14
- 4 The strongest corner joint for drawer construction is the half lap miter dado dovetail
- 5 Oak is fumed with sulphuric acid wood alcohol acetic acid ammonia

¹ From Test 5 of Ruch-Cossmann Biology Test: Form A (World Book Company).

The following 5-response test illustrates a variation of the method for the sake of economy of scoring. The numbering of the responses permits the most rapid scoring of any test technique yet devised. The first use of numbered responses seems to have been in some clerical tests for silk workers designed by Dr. A. S. Otis.

II. A 5-Response Test in English Literature¹

DIRECTIONS. Read each question and select the *best* answer to that question.

Record the *number* of the best answer on the dotted line, as shown in the following samples. (Samples omitted here.)

¹ From the Iowa High School Content Examination: Form A, Part I (Extension Division, University of Iowa, Iowa City).

TRUE-FALSE TESTS

DIRECTIONS. If a statement is true, underline *True*; if false, underline *False*. If in doubt, omit the item. *Do not guess.*

1 One centimeter is a little more than 2.54 inches.	<i>True</i> <i>False</i>
2 An object immersed in water is buoyed up with a force which is numerically equal to the weight of the water displaced.	<i>True</i> <i>False</i>
3 When an ordinary electric-light bulb is broken, the glass flies away from the center of the bulb in all directions.	<i>True</i> <i>False</i>
4 1000 cc. of water is approximately one pint.	<i>True</i> <i>False</i>
5 The "kick" of a rifle when fired may be explained by Newton's third law of motion.	<i>True</i> <i>False</i>

Some test users have objected on *a priori* grounds to the use of true-false statements in the form just given. They hold that the presentation of wrong statements to the pupil will fix error in his mind. To avoid this, such objectors advocate the use of the question form of statement, thus:

Did Columbus discover America in the year 1895? *Yes* *No*
(or *True* *False*)

Although there is no experimental evidence which suggests the necessity of the question form, there can be no valid objection to its use in order to "play safe." From a psychological point of view, it seems unlikely that false teaching can arise from either form of statement, since the directions always state that some of the statements are true and some are false. This gives the child a critical mental set or attitude toward the test, and probably reduces to a minimum any danger of false teaching. Life presents its statements and situations in true, false, and mixed form, and hence there are certain very strong arguments for the social utility of the true-false method.

MATCHING EXERCISES

DIRECTIONS. Read each *characterizing phrase* and then find the *man* at the left whom the phrase fits best. Record the *number* of the proper man in the parentheses in front of each phrase.

MEN	CHARACTERIZING PHRASE
1 Thomas H. Benton	(5) Author of the Declaration of Independence
2 Thaddeus Stevens	() For thirty years a senator from Missouri
3 George B. McClellan	() An immigrant who worked for political reform
4 Carl Schurz	() Leader of the Union Army in Peninsula Campaign
5 Thomas Jefferson	() Congressman demanding harsh treatment of South
6 Miles Standish	() Discoverer of the New World for Spain
7 De Witt Clinton	() Spent a fortune to found a colony in America
8 Charles Sumner	() Military man of Plymouth, told of by Longfellow
9 Sir Walter Raleigh	() Massachusetts senator denouncing the "Crime against Kansas"
10 Christopher Columbus	() Governor of New York, promoted the Erie Canal

The following is a portion of a matching exercise over certain terms in European history. The extracts cannot be completely matched, as reproduced, due to omissions. The directions are similar to those of the foregoing matching test. This test, however, would seem to call for considerably more thought and comprehension than the preceding one, and hence is probably a better type of examination.

Characteristics of Objective Examinations 277

TERMS	STATEMENT OF SIGNIFICANCE
1 Submerged nationality	() The passing of the tariff and other laws favorable to a nation's industrial and commercial development
2 Sphere of influence	() All acts on the part of employees willfully to reduce the output of industries, the purpose being to terrorize owners into meeting the demands of workers
3 "Open door" policy	() The principle of referring to the people of a nation important legislative matters
	* * * * *
6 Protectionism	() A proposal of the United States that the citizens of all nations should have equal rights for commercial and industrial advantages in China and that all nations respect China's integrity
7 Initiative	() A group of homogenous people cherishing the sentiment of becoming a nation, but who are ruled by a nation of a different nationality which for selfish reasons does not wish to grant the group its freedom; as the Poles, Slavs, Italians, etc., under the rule of Austria
	* * * * *
15 Labor Union	() A movement in Italy which is opposed to a violent overturn of the usual methods of business and manufacturing; accused of suppressing liberty and of checking the communists by violence; has done much to do away with strikes and to re-establish conditions as they were before the World War

BEST ANSWER OR JUDGMENT TESTS

DIRECTIONS. Below are a number of incomplete statements which may be completed by any one of three possible answers. Only one answer is scientifically correct; the other two are partly or entirely incorrect. Study each statement and then make a cross in front of the *best* answer, as shown in the sample. (Sample omitted here.)¹

- 1 The chief function of the white blood corpuscles is the
— Destruction of disease germs in the blood.
— Carrying of oxygen to the tissues.
— Carrying of food materials to the cells of the body.
- 5 The best of these definitions of photosynthesis is
— The action of sunlight on plants.
— The process of food manufacture in green plants.
— The process by which plants give off oxygen.
- 9 A vein is best defined as a blood vessel carrying
— The blood going to the heart.
— The "blue blood."
— The "impure blood."
- 16 The chief advantage of transpiration to plants is thought to be
— Elimination of nitrogenous waste materials.
— Cooling the plant on very hot days.
— Insuring the proper absorption of minerals.

The following are a selected group of items from the examination in woodworking previously quoted from :

- 7 Three-ply panels are better than solid panels because
— They take a better finish.
— They cost less.
— They are less apt to check or warp.
- 13 Hot glue is preferred over cold glue because
— It costs less.
— It penetrates the pores of the wood better.
— It is more easily applied.
- 14 Oil stains should not be used on first-class work because
— They injure the wood.
— They are expensive.
— They do not give clear effects.

¹ Ruch-Cossmann Biology Test: Form A, Test 2.

It will be seen that these so-called best-answer or judgment tests are in reality multiple-response tests of a somewhat different mechanical form. The chief advantage of the best-answer type lies in the longer statements possible, greater ease of reading, and somewhat greater usefulness for reasoning purposes. They are extravagant of space, however. Such tests would seem to have wide applicability and cannot be objected to on the basis of unnaturalness to any great extent.

A somewhat different use of the same principles will be found in the following exercise, which has certain advantages in the extent to which chance affects results.

DIRECTIONS. Check (✓) the statement which expresses what you might have prophesied as to the future of the Roman Republic, if you had lived during the first century before Christ and had known the following facts:

Marius becomes consul for the seventh time;
Sulla is given the title of "Perpetual Dictator";
Cæsar becomes dictator for life.

- The republic was on the verge of developing a greater democracy.
- The army becomes less aristocratic, and Marius enlists all men who desire to fight.
- The senate desired to grant great economic rights to the working class of Rome.
- Civil wars and the military rule of one-man power would in time overthrow the republic.
- The rule of the assembly and its leaders was about to triumph over the rule of the senate.

(Etc., to any practicable number of statements.)

IDENTIFICATION EXERCISES

DIRECTIONS. Select the *best* breakfast in the four breakfasts given below for a boy or girl twelve years old. In making your choice, use all the facts that you have just learned in your study of balanced meals.

Answer. My first choice as the best breakfast is No.

BREAKFAST No. 1

$\frac{3}{4}$ cup of oatmeal, cream and sugar
2 eggs and fried potatoes
3 slices of bread and butter
1 cup of coffee

BREAKFAST No. 2

1 shredded wheat biscuit
2 slices of toasted brown bread with butter
1 dish of prunes
1 glass of milk

BREAKFAST No. 3

1 dish of cornflakes
2 eggs, with 2 pieces of toast
3 ginger cookies
1 cup of coffee

BREAKFAST No. 4

3 buckwheat cakes with honey
1 soft-boiled egg
3 slices of toast with currant jelly
1 cup of tea, cream and sugar

The above exercise¹ is perhaps better adapted for teaching purposes than tests and examinations proper. It would be expensive to mimeograph. Such an exercise, if written on the blackboard, would be an excellent basis for a 15-minute class discussion of several fundamental principles of dietetics.

REARRANGEMENT TESTS

DIRECTIONS. Number the following great inventions in the order of their making. Put 1 before the earliest, 2 before the next in order, and continue for the others, giving the most recent invention the number 10.

..... The printing press The cotton gin
..... Gunpowder The automobile
..... Wireless telegraphy The steam engine
..... Engraving Arabic numerals
..... The telephone The mariner's compass
(Etc., to ten)	(Etc., to ten)

¹ Quoted, with changes, from Ruch, G. M., *The Improvement of the Written Examination* (Scott, Foresman & Co., 1924), pages 72-73.

The foregoing examples do scant justice to the many variate types of objective examinations which have been developed to date. Many school subjects have not been illustrated at all. Some of the individual items among these illustrations cannot be defended for unambiguity and accuracy. However, with a few exceptions, they are actual examinations built by classroom teachers, and most of them have actually been tried out in the classroom.

For a wide variety of objective examination types, see the volumes by Monroe, Paterson, Ruch, and Russell cited in the bibliography at the end of Part III.

CHAPTER SIXTEEN

CRITICAL CONSIDERATIONS IN OBJECTIVE EXAMINATION METHODS

Introduction. There are a large number of fundamental issues in the administration and scoring of objective tests upon which experimentation is greatly needed. Two of these issues which have attracted attention are: (a) the validity of the conventional formulas for correcting for chance or guessing in multiple-response tests, and (b) whether pupils should be instructed to omit doubtful items or to attempt every one. Practices vary on both these points, although the experimental evidence seems now to point toward tentative decisions which can be put into practice pending future work.

Corrections for chance or guessing. In true-false and other recognition types of tests, where several alternative responses are presented, it has been the practice to make statistical allowances for the effects of chance, at least when the number of responses is three or fewer. The general formula used may be stated:

$Score = Rights - \frac{Wrongs}{n - 1}$; where n is the number of responses presented, and from which the selection of the one correct response is to be made. This formula can also be written in more specific form, as follows:

(For 2-response and true-false

tests) Score = Rights - Wrongs

(For 3-response tests) . . . Score = Rights - $\frac{1}{2}$ Wrongs

(For 4-response tests) . . . Score = Rights - $\frac{1}{3}$ Wrongs

Etc.

The use of the chance correction formula may be shown by two examples, a true-false test (mechanically a 2-response test) and a 3-response test being selected.

TRUE-FALSE		3-RESPONSE	
Total number of items	100	Total number of items	150
Omissions	3	Omissions	15
No. attempted (<i>Subtract</i>)	97	No. attempted (<i>Subtract</i>)	135
Number wrong	13	Number wrong	30
Number right (<i>Subtract</i>)	84	Number right (<i>Subtract</i>)	105
Number wrong	13	$\frac{1}{2}$ the wrongs ($30 \div 2$)	15
Rights—wrongs (<i>Subtract</i>)	71 (Score)	Rights— $\frac{1}{2}$ wrongs (<i>Subtract</i>)	90 (Score)

It should be noted that in the case of the true-false tests the corrected score also equals (*Attempts*)—(2 times the *Wrongs*). This has led some teachers to regard the latter method in the light of a double penalty for errors. The more detailed explanation above shows the fallacy in viewing the correction technique as a penalty. In a 100-question true-false examination, a pupil possessing zero knowledge about all 100 questions, if forced to guess at all questions, would be expected to earn approximately 50 correct answers. The *Rights minus the Wrongs* would then equal roughly 50–50 or 0. The correction method is therefore valid within the limits which a theorem in probability may be expected to hold with relatively small numbers of questions.

There is another aspect of this question which may be discussed in passing; viz., to what extent pupils actually guess at answers. There are at least hypothetical grounds for believing that pure guessing in a test is partly a matter of temperament on the part of the pupil. It is difficult both to force certain pupils into guessing and also to restrain others from guessing when thought would lead to the right answer. Again, there is the matter of misinformation. In the item,

“Gold is more dense than platinum . . . *True False*,” the pupil might underline *True* not because it is a 50-50 break but because he actually thinks the statement is true. It is doubtful if such cases of positive misinformation iron out in the process of correction.

Like the foregoing statements, much of the literature on the subject of chance effects in multiple-response tests is highly speculative. The question of the validity of the chance formula can be decided only by experimentation, and the further discussion of this point will be deferred until certain experiments have been summarized.

Instructions to guess or not to guess. Various writers have taken issue on this matter, but as far as the present authors can learn there has been but one experimental investigation of the effectiveness of verbal instructions about guessing at doubtful items.¹ McCall² has held that pupils should be told to guess in such cases. Wood³ has systematically favored the "do not guess" instructions.

In addition to the weight of certain investigational evidence to be brought out later, Wood's position would seem to have the advantage in the general logic of the situation. There would seem to be no very good reason for forcing guessing to the maximum in order that a statistical correction might work out better. Most teachers would probably agree that it is safer not to encourage guessing at doubtful items.

Experimental data on guessing and corrections for chance. The most extensive investigation to date which covers both the effects of instructions about guessing and also the effects upon the validity and reliability of test scores arising from the use of the conventional correction formula seems to be that of DeGraff and Ruch.⁴ A total of 2453 pupils was used in this study, these being broken into 10 groups of approximately equal ability.

¹ Ruch, G. M., et al., *Objective Examination Methods in the Social Studies* (Scott, Foresman & Co., 1926).

² *Journal of Educational Research*, Vol. I (1920), pages 33-46.

³ Wood, Ben D., *Measurement in Higher Education* (World Book Company, 1923), pages 219, 232, 251, 257, and elsewhere.

⁴ Ruch, G. M., et al., *Objective Examination Methods in the Social Studies* (Scott, Foresman & Co., 1926), Chapter IV, pages 54-88.

Two forms of a recall or completion test of 100 items each were prepared and given on two consecutive days to 2453 pupils in Grades VII, VIII, XI, and XII. These two forms proved to be approximately of the same average difficulty. They will be designated as Forms A and B.

The same items (as nearly as this could be accomplished and still be branded as "same") were then "translated" into 7-response, 5-response, 3-response, 2-response, and true-false forms, in turn, keeping the same allotment of items to the A and B forms. It is true that more or less distortion of the original recall or completion items must have taken place during such a process, but the following selected items will show in general the sort of changes which were necessary.

Recall or Completion

1 Christopher Columbus discovered America in the year
50 President Wilson said, "The world must be made safe for
100 The man who invented the process of hardening rubber was.....

7-Response

1 Christopher Columbus discovered America in the year —
(1) 1498 (2) 1492 (3) 1607 (4) 1450 (5) 1619
(6) 1592 (7) 1540

50 President Wilson said, "The world must be made safe for —
(1) private property (2) autocracy (3) socialism
(4) plutocracy (5) labor (6) aristocracy (7) democracy

100 The man who invented the process of hardening rubber was —
(1) Edison (2) Whitney (3) Ford (4) McCormick
(5) Field (6) Goodyear (7) Bell

5-Response

(Two of the wrong responses were dropped; otherwise no changes.)

3-Response

(An additional two wrong responses were dropped; otherwise no changes.)

2-Response

1 Christopher Columbus discovered America in the year —
 (1) 1498 (2) 1492

50 President Wilson said, "The world must be made safe for —
 (1) autocracy (2) democracy

100 The man who invented the process of hardening rubber was —
 (1) Edison (2) Goodyear

True-False

1 Christopher Columbus discovered America in the year 1492. *True False*

50 President Wilson said, "The world must be made safe for autocracy." *True False*

100 The man who invented the process of hardening rubber was Edison. *True False*

This process gave five variant editions (7-, 5-, 3-, 2-response and true-false) in addition to the recall arrangement. In these five editions both Forms A and B were printed as one booklet (i.e., all 200 items, Form B following Form A).

The next step was the matter of writing the instructions to the pupils. Half of the printed booklets for each of the five editions just described were prepared with instructions for the pupils to guess at all doubtful items, and half were prepared with directions not to guess but to leave all doubtful items unanswered. The exact wording of these two sets of instructions is reproduced here, the 5-response version being chosen as the illustration.

“Guess” Edition

NOTE CAREFULLY: Do not leave any questions unanswered. If you don't know, guess. It is better to guess than to leave a question blank, because you have one chance in five of getting it right by pure guessing. You should try to make as logical or shrewd a guess as possible.

REMEMBER: Try to answer every question. Guess if you do not know.

“Do Not Guess” Edition

NOTE CAREFULLY: If you are in doubt about the answer to any question, leave it blank. Do not guess! You will be penalized for all wrong answers. The tests are scored in such a way that you will lose more than you will gain by guessing.

REMEMBER: Do not guess. Answer only those that you are reasonably sure about.

Considering the five multiple-response versions and the two editions relative to instructions about guessing, this made a total of 10 different variations of the tests (both Forms A and B) in addition to the Recall, Form A, and Recall, Form B.

The entire plan of the experiment can now be set forth in outline form. Three test sittings were planned and carried out, as follows:

- I. First Sitting (first day): Recall, Form A; all 2453 pupils;
- II. Second Sitting (second day): Recall, Form B; all 2453 pupils;
- III. Third Sitting (third day): the 2453 pupils were given the 10 variant editions of the multiple-response (and true-false) tests in scrambled fashion so that this total group was broken by chance into 10 sub-groups.

The plan of the third sitting insured that approximately equal numbers (actually 229 to 281) of pupils took each of the

10 variant editions of the tests. Also, the procedure guaranteed the rough equivalence of ability in the 10 sub-groups, since the 2453 blanks were distributed totally at random, each teacher dealing her allotment of blanks from the top of the pile to the pupils in the order of seating in the room. Several dozen schools and many different states and courses of study were thus sampled by this method of chance distribution.

Since all pupils took both forms of the recall tests, it was possible to use the scores on the recall examinations as a criterion of the pupils' actual knowledge of these 200 facts from United States history. It should be noted that the recall version was selected as the criterion because it was the freest from the factors of chance and guessing.

Several lines of statistical analysis of the results were made possible by the arrangement of the experiment. A few matters by way of definition and explanation of abbreviations are needed in connection with Tables 65, 66, and 67, which follow.

The term "validity coefficients" always refers to correlations of the scores earned by pupils on the recall test with one of the other 10 variations of the test. The "reliability coefficients" are the correlations of Forms A and B for the same variate of the test. "Guess" or (G) refers to the edition suggesting that pupils guess at all doubtful items. "Do not guess" or (N) refers to the alternate editions forbidding guessing. "Unc." means scores to which the chance correction formula ($\text{Rights} - \text{Wrongs}/(n - 1)$) has not been applied. "Corr." refers to the test scores after correction by this formula. The italicized differences are those that are statistically significant — i.e., three or more times as large as their probable errors. In other words, the chances that the obtained differences are not due to chance alone are about 20:1.

Table 67 is given in order to show how equal the sub-groups proved to be in ability. The differences between sub-groups are probably small enough to be ignored in the practical interpretation of the results.

TABLE 65
VALIDITY COEFFICIENTS

RECALL <i>rs.</i>	"GUESS"			"DO NOT GUESS"			DIFFERENCES			
	(1) Unc.	(2) Corr.	(3) Diff. (2-1)	(4) Unc.	(5) Corr.	(6) Diff. (5-4)	(4-1)	(4-2)	(5-1)	(5-2)
7-Response A	.871	.873	.002	.927	.926	-.001	.056	.054	.055	.053
7-Response B	.816	.861	.045	.872	.898	-.026	.056	.011	.082	.037
5-Response A	.907	.910	.003	.891	.918	.027	-.016	-.019	.011	.008
5-Response B	.860	.903	.043	.836	.870	.034	-.024	-.067	.010	-.033
3-Response A	.838	.848	.010	.845	.915	.070	.007	-.003	.077	.067
3-Response B	.797	.875	.078	.852	.902	.050	.055	-.022	.105	.027
2-Response A	.859	.865	.006	.740	.775	.035	-.119	-.125	-.084	-.090
2-Response B	.735	.806	.071	.752	.868	.116	.017	-.054	.133	.063
True-False A	.804	.839	.035	.749	.860	.111	-.055	-.090	.056	.021
True-False B	.675	.801	.126	.768	.856	.088	.095	-.033	.181	.055
Mean <i>r</i>	.815	.858		.823	.890					
Proportion of "significant" differences (italicized values) to total number of differences (10)	+ - Both	2:10 0:10 2:10			5:10 0:10 5:10	2:10 1:10 3:10	1:10 3:10 4:10	7:10 1:10 8:10	3:10 1:10 4:10	

TABLE 66
RELIABILITY COEFFICIENTS

TEST	"GUESS"			"DO NOT GUESS"			DIFFERENCES			
	(1) Unc.	(2) Corr.	(3) Diff. (2-1)	(4) Unc.	(5) Corr.	(6) Diff. (5-4)	(4-1)	(4-2)	(5-1)	(5-2)
Recall (.950)										
7-Response	.800	.839	.039	.886	.907	.021	.086	.047	.107	.066
5-Response	.864	.902	.038	.862	.882	.020	-.002	-.040	.018	-.020
3-Response	.837	.858	.021	.886	.890	.004	.049	.028	.053	.032
2-Response	.745	.864	.119	.859	.843	-.016	.114	-.005	.098	-.021
True-False	.641	.780	.139	.885	.837	-.043	.241	.105	.196	.059
Mean <i>r</i>	.777	.849		.876	.872					

NOTE. The italicized values show all differences which are 3.0 or more times their probable errors, and hence are probably "significant differences."

TABLE 67

MEAN SCORES

	RECALL	MULTIPLE-RESPONSE		RECALL	MULTIPLE-RESPONSE			
		A			B	B		
		Unc.	Corr.			Unc.	Corr.	
7-Response (G) ¹	25.9	50.0	41.5	26.2	39.6	32.6		
7-Response (N) ²	27.6	44.9	40.0	27.6	37.2	33.1		
5-Response (G)	25.7	54.2	43.4	26.9	45.5	35.4		
5-Response (N)	28.0	48.8	42.3	28.6	42.1	36.4		
3-Response (G)	25.6	62.2	43.6	26.1	55.5	36.6		
3-Response (N)	27.4	54.1	41.9	27.5	48.2	36.1		
2-Response (G)	26.7	71.7	43.6	27.4	67.2	37.1		
2-Response (N)	33.4	65.1	45.8	33.3	60.3	40.2		
True-False (G)	27.4	65.8	32.3	26.8	61.3	26.0		
True-False (N)	27.5	51.0	30.8	27.6	47.6	26.8		

¹(G) means instructions to "Guess."

²(N) means instructions "Do Not Guess."

Conclusions from the preceding investigation. The three tables immediately preceding lead to the following tentative conclusions:

1. The evidence on the whole points toward slight advantages in favor of instructing pupils not to guess, but to omit doubtful items. Instructions not to guess seem to have some effectiveness, as the average scores are thereby lowered. The validity of do-not-guess instructions is slightly higher, possibly due to the fact that such instructions appear to raise the reliability of the results.
2. Instructions not to guess may be somewhat more necessary when no corrections for chance are employed.
3. The use of the chance correction formula measurably increased the validity of the results under both sets of instructions.
4. The best technique for administering multiple-response tests would seem to be the combined use of (a) instructions not to guess and (b) corrections for chance.
5. The effects of both factors under study seem to be most evident in case a small number of responses are used; i.e., in the 3-response, 2-response, and true-false tests.
6. For practical purposes, the differences in the various techniques are not large enough to rule out any particular combination of methods completely. Both the validity and reliability can be raised easily by the expedient of longer tests. It is probably true that 150 items under "guess" instructions and without correction are at least as valid as 100 items given under "do not guess" directions and with corrections for chance.
7. The practice of instructing pupils to guess appears to have no advantages over the reverse instructions, and it does not insure a better "working out" of the formula for correcting for chance.

8. There are certain indications in the behavior of the average scores for true-false tests in comparison with 2-response tests, both with and without correction, which raise the question whether true-false tests are 2-response tests in any more real sense than their mechanical arrangement.

9. All forms of the recognition tests, both before and after correction, proved markedly easier than the recall type.

10. After correction for chance, all the average scores on the recognition tests except the true-false are roughly equal; the true-false tests when corrected are very much more difficult than the other recognition forms, and, in fact, they are almost as difficult as the recall. This suggests that if the corrections are approximately just in the case of the multiple-response tests, the true-false tests are *over-corrected* by the formula. If the true-false tests are held to be properly corrected for chance, the multiple-response tests appear to be *under-corrected*.

Results from similar studies. The published experimental work on the effects of chance corrections is somewhat confusing. The authors¹ published the results of two minor investigations which seemed to show that the application of the chance correction formula lowered the reliability of test scores. This finding was confirmed by Paterson and Langlie,² but the later work of Wood³ and Ruch⁴ from the standpoint of validity rather than reliability leads to conclusions more in harmony with the ten points summarized in the preceding section of this chapter.

For the present, therefore, we seem to be on safe grounds to combine the use of instructions against guessing with the use of the correction formula ($Score = Rights - Wrongs / (n - 1)$), where n is the number of responses presented.

¹ *Journal of Educational Psychology* (February, 1925), pages 89-103.

² *Journal of Applied Psychology* (December, 1925), pages 339-348.

³ *Journal of Educational Psychology* (January-April, 1926).

⁴ Reference cited in the preceding section.

Time allowances for objective tests. It is of course impossible to give more than a rough sort of guide to the number of items which a pupil can answer in a stated length of time. Much depends upon the school subject, upon the pupil's normal speed of work, whether the tests are chiefly factual or reasoning materials, and the degree of difficulty of the examination as a whole.

Basing the following statements on the study which was just summarized, the average high school pupil works at the rates given below for difficult factual examinations.

TYPE OF TEST	RATE OF ANSWERING
Recall or Completion	About 4 items per minute
7-Response	About 4 items per minute
5-Response	About 5 items per minute
3-Response	About 5 items per minute
2-Response	About 6 items per minute
True-False	About 6 items per minute

Many records have, however, shown rates of work at least double those given above for fairly easy and highly factual types of examinations. When the test is intended to measure thought and reasoning, the numbers of items per unit of working time must be greatly decreased over the figures suggested here.

The greater rapidity with which true-false, 2-response, and 3-response tests can be answered allows such tests to be made longer in terms of numbers of items in comparison with many-response and completion tests. This fact alone will explain why the few-response tests *per unit of working time* give just about as satisfactory results as do the many-response tests which are less influenced by chance factors.

Summary and conclusions. In concluding this treatment of informal objective examination methods, several statements should be made. In the first place, the discussions in Part III of this volume are at best very sketchy and they

pass over lightly or ignore many interesting questions for reasons of limitations of space and the highly controversial literature covering much of this field. Accordingly, the references at the end of the chapter should be consulted by the reader who is interested in the critical study of objective examination techniques.

Secondly, it should not be overlooked that the data of Chapter XVI, in comparison with those of Chapter XIV, point very clearly to the superiority of objective examinations over the traditional test in terms of the degree of validity and reliability attainable per unit of examination time. The objective examination is, to be sure, in its infancy, but it is indeed a promising child.

Lastly, as was mentioned in an earlier section, Part III might give the casual reader the impression that no legitimate place is accorded the traditional written examination in the program of instruction and measurement in the high school. This was far from the intention of the authors; their attitude is that where personal opinion can be eliminated, it should be. There is ample room for the coexistence of the standard test, the ordinary written examination, and the informal objective examination. These serve to supplement one another and to provide a system of checks and balances one against the other. The greatest usefulness of each does depend, however, upon its application and interpretation, with full knowledge of the inherent advantages and limitations of the method.

Selected References for Part III

I. Books and Monographs

MONROE, W. S., and SOUDERS, L. B. *Present Status of Written Examinations and Suggestions for Their Improvement*. Bulletin No. 17 (1923). Bureau of Educational Research, University of Illinois, Urbana, Illinois. (Presents experimental studies of the reliabilities of written examinations and samples of objective examinations.)

PATERSON, D. G. *Preparation and Use of New-Type Examinations*. World Book Company, Yonkers-on-Hudson, New York; 1925. 87 pages. (Discussions of the merits and limitations of the various types of objective examinations, chiefly on the college level. This volume is very suggestive, however, of test techniques useful in the high school. Contains an excellent bibliography of more than fifty titles.)

RUCH, G. M. *The Improvement of the Written Examination*. Scott, Foresman & Co., Chicago; 1924. 193 pages. (The first volume devoted exclusively to objective examination methods. Discusses the functions, criteria, sources of error, types, experimental studies, and statistical considerations related to both old and new types of examinations. About 60 pages of this book are devoted to illustrative examinations of the objective variety. Adapted to both elementary and high schools.)

— et al. *Objective Examination Methods in the Social Studies*. Scott, Foresman & Co., Chicago; 1926. 126 pages. (A report of an investigation of examinations in the social studies under a grant from the New York Commonwealth Fund. Devoted entirely to experimental studies of the reliability of official eighth-grade examinations, the validity of corrections for chance, the merits of "guess" instructions versus "do not guess" directions, and the comparative validities and reliabilities of recall, multiple-response, true-false, and matching tests. Somewhat more technical than the other references cited above. The most comprehensive study yet published on several debated issues in examination techniques.)

RUSSELL, C. *Classroom Tests*. Ginn & Co., Boston; 1926. 346 pages. (In two parts — Part I discusses why and how to make classroom tests; Part II tells why and how to use classroom tests. The sample examinations given are chiefly for elementary school subjects, but the general discussions are applicable on any level. The practical advice is unusually good. A great deal of space is devoted to the treatment of test scores. The value of the book is lessened somewhat by the failure of the author to utilize the experimental work which has been done in the field of examination construction.)

WOOD, BEN D. *Measurement in Higher Education*. World Book Company, Yonkers-on-Hudson, New York; 1923. 337 pages. (Although this volume is chiefly devoted to intelligence testing on the college level, it is nevertheless a very valuable contribution to the technique of the

newer types of examinations. Chapters VII and VIII contain one of the best discussions of the theory and principles of examinations which has appeared in print. Chapters IX to XII present a wide variety of objective tests over college subjects. The critical attitude of the author is especially commendable.)

II. Articles

This section of the References is purposely restricted to a selected list of titles, preference being given to experimental studies, since the more introductory and general discussions have been taken care of in the larger treatments cited above.

ASKER, W. "The Reliability of Tests Requiring Alternative Responses." *Journal of Educational Research*, Vol. IX (1924), pages 234-241. (Advocates, among other things, the discouragement of guessing.)

CHAPMAN, J. C. "Individual Injustice and Guessing in the True-False Examination." *Journal of Applied Psychology*, Vol. VI (1922), pages 342-348. (Holds that the *R-W* method of correcting true-false tests does not neutralize guessing effects.)

— and TOOPS, H. A. "A Written Trade Test: Multiple-Choice Method." *Journal of Applied Psychology*, Vol. III (1919), pages 358-365. (One of the earlier papers which will prove suggestive to teachers of the manual arts.)

FILER, H. A., and O'ROURKE, L. J. *Annual Reports of the Chief Examiner and the Director of Research of the U. S. Civil Service Commission for the Fiscal Year Ended June 30, 1923*. Government Printing Office, Washington, D. C.; 1923. (Statistical and descriptive account of objective examination methods tried out in civil service examinations.)

HAHN, H. H. "A Criticism of Tests Requiring Alternative Responses." *Journal of Educational Research*, Vol. VI (1922), pages 236-240. (Opposes the *R-W* method of scoring and attacks the pedagogical desirability of true-false tests.)

KOHS, S. C. "High Test Scores Obtained by Subaverage Minds." *Psychological Bulletin*, Vol. XVII (1920), pages 1-5. (Shows the slight probabilities of high scores by chance alone on 50-item tests.)

ODELL, C. W. "Another Criticism of Tests Requiring Alternative Responses." *Journal of Educational Research*, Vol. VII (1923), pages 326-330. (A defense of the correction formula for chance.)

PATERSON, D. G., and LANGLIE, T. A. "Empirical Data on the Scoring of True-False Tests." *Journal of Applied Psychology*, Vol. IX (1925), pages 339-348. (The chance correction is shown to lower the reliability of test scores.)

RICHARDS, O. W., and KOHS, S. C. "High Test Scores Attained by Subaverage Minds." *Journal of Educational Psychology*, Vol. XVI (1925),

pages 8-18. (Presents arguments from theorems of probability for making true-false tests at least 75 items in length.)

RUCH, G. M., and STODDARD, GEORGE D. "Comparative Reliabilities of Five Types of Objective Examinations." *Journal of Educational Psychology*, Vol. XVI (1925), pages 89-103. (Data on reliabilities of recall, multiple-choice, and true-false tests with and without correction for chance. Chance corrections seemed to lower reliability. In this connection, however, see the second reference under Ruch in Section I above and the articles cited for Wood below. The later experiments of Wood and Ruch seem to show that validities are raised by chance corrections in spite of the fact that reliabilities may be lowered.)

TOOPS, H. A. "Trade Tests in Education." *Teachers College Contributions to Education*, No. 115 (1921). Columbia University, New York. (One of the earliest experimental studies of the comparative merits of recall, multiple-response, and true-false tests. Gives data on time limits and comparative reliabilities per equal time allowances.)

WOOD, BEN D. "Studies of Achievement Tests." *Journal of Educational Psychology*, Vol. XVII (1926), pages 1-22, 125-139, 263-269. (Agrees with Ruch and Stoddard and with Paterson and Langlie that chance corrections lower reliability. Advocates chance corrections on basis that validities are raised (conclusion reached also by Ruch et al in later investigations). Perhaps the best critical discussion of various technical points in test construction yet in print. Report of a study financed by the New York Commonwealth Fund.)

PART FOUR

THE CONSTRUCTION OF EDUCATIONAL AND MENTAL TESTS

CHAPTER SEVENTEEN

THE CONSTRUCTION OF EDUCATIONAL AND MENTAL TESTS

I. VALIDATION

Introduction. No single formulation can be made covering the detailed stages in the construction of a test or scale. Mental tests present many special features in their derivation not common to educational tests. The reverse is also true. *Tests*, proper, in contrast with *scales*, involve important differences in technique. The rules which might be laid down for the building of a test in an elementary school subject would not serve equally well for secondary subjects. There are also unique principles, not common to school tests, which have been found useful in the development of trade tests, prognosis tests, and tests of many psychological capacities like musical talent, will-temperament, mechanical aptitudes, etc.

Chapters XVII to XX will present in considerable detail the sequence of steps in the production of typical tests. No particular attempt will be made to confine the treatment to high school tests, and variations for special applications will be pointed out from time to time. The following general outline will serve as a skeleton for the description of the several stages in the standardization of a test.

OUTLINE OF THE CONSTRUCTION OF TESTS AND SCALES

I. Validation of the test

1. Setting up the *criterion* of validity
2. Original selection of items
3. Experimental try-out of items
4. Computation of difficulties of items
5. Scaling or weighting the test items

- II. Breaking the test into equivalent forms
 - 1. Assignment of items to the new forms
 - 2. The second try-out for equivalence of forms
 - 3. Final determination of time limits
- III. Derivation of norms
 - 1. Third or final try-out for derivation of norms
- IV. Determination of reliability of the test
 - 1. Calculation of reliability coefficients
 - 2. Calculation of measures of error in individual scores
- V. Perfecting the administration of the test
 - 1. The Manual of Directions
 - 2. Answers — keys or stencils

SETTING UP A CRITERION OF VALIDITY

The meaning of validity. Validity has previously been defined as the degree to which a test or scale measures what it is claimed to measure. Validity is not a purely statistical concept like reliability, but includes a wide variety of elements which must be handled separately in the validation of the test. In a general way, the validation of a test consists in the selection of test items of prime importance and in the elimination of unimportant items or those proving erratic in behavior under experimental try-out. General validity under such definitions as given here cannot exist. Tests must be validated for specific purposes; and regardless of the degree of validity for such purposes, the validity cannot be guaranteed if the test is wrongly administered or misapplied.

The criterion or criteria of validity. The setting up of adequate criteria for the selection of test items is the first, most important, and most difficult step in the building of a test. Such a criterion is necessarily different with each individual test or scale. Some notion of the general meaning of the expression "setting up a criterion" may be had from

the descriptions given by Terman for the validation of the Stanford Revision of the Binet-Simon Scale.¹ A number of different principles were adopted as criteria of the merit of individual test items, somewhat as follows:

- (1) The test items should be relatively free from the influence of schooling; i.e., not be taught specifically during the years of formal education.
- (2) The information required for success on a test must be common to the experience of all children; the use or application of this information must, however, be *novel*.
- (3) The tests must not be measures of sensory capacities (little related to intelligence) but should call into play the "higher" mental processes of judgment, discrimination, analysis, synthesis, reasoning, etc.
- (4) Tests which show evidence of *spontaneous* interest in the common environmental phenomena are probably measures of mental alertness or intelligence.
- (5) The individual tests must show a steadily increasing percentage of successful responses at each successive age level if applied to unselected age groups (as an evidence that the tests reflect differences in mental maturity).
- (6) Each test should show a higher percentage of successes for children of a given age who are *known* to be bright than for those who are known to be dull.
- (7) The tests should *sample widely* over a range of abilities, talents, and capacities.
- (8) The scoring of the items must be as objective as possible; i.e., be as free as possible from the personal equation of the examiner.

¹ Terman, L. M., et al., *The Stanford Revision and Extension of the Binet-Simon Measuring Scale of Intelligence* (Warwick & York, Inc., 1916). Terman, L. M., *The Measurement of Intelligence* (Houghton Mifflin Company, 1916).

The formulation of criteria of the validity of educational test items requires a different set of principles in many ways, although the general spirit of the procedure is the same. In either case *the prime consideration is the delineation of a set of specific guiding objectives and conditions which the test items must meet to be held valid.* It should be kept in mind that the reasoning here is perfectly analogous to that of curriculum construction, and the validation of a test can be no better than the present state of knowledge about the objectives, aims, minimum essentials, social utility, etc., of the curricular content.

ORIGINAL SELECTION OF ITEMS

Methods. Study of the published accounts of the validation of existing tests shows a wide variety of methods. The more frequently used validation methods are:

- (1) Textbook analysis
- (2) Analysis of courses of study
- (3) Analysis of final examination questions
- (4) Pooled judgments of competent persons
- (5) Use of rating scales in setting up criteria
- (6) Correlations with school marks or other measures of school success
- (7) Increase in percentage of successes with successive ages or grades
- (8) Correlations with previously validated measures
- (9) Differential scores shown by two groups known to be widely separated upon a scale of ability
- (10) Determination of social utility
- (11) Logical or psychological analysis
- (12) Correlations with tests of other intellectual, non-intellectual, or educational abilities

1. *Textbook Analysis*

Validation by textbook analysis. Many pertinent objections may be raised against the validation of test items by statistical analysis and summary of textbooks. There are also important advantages of this method. The objections will be considered first.

A test which represents nothing more than a composite picture of the content of ten, fifteen, twenty, or more representative textbooks in a given school subject cannot rise above the level of measuring *what is actually taught* in the rank and file of schools. Such a test fails to a degree, because it does not measure *what ought to be taught*. It represents a static content, and it is not as likely to stimulate beneficial changes in teaching methods as a test based primarily upon what is socially useful.

Against these disadvantages may be set a number of favorable considerations. In subjects which are yet in the developmental stage, such as general science, correlated mathematics, and the social studies, this method is practically the only one open to the test builder at the present time. Tests thus constructed do tend to fit rather closely the teaching practices of the day. A series of test items obtained by the criterion of occurrence in from 50 to 100 per cent of twenty-five leading textbooks in a given field certainly can be said to be the *most probably representative* of general teaching practice in that subject. Such items are valid under normal or typical conditions. Such a test may fail to do justice to the classes of that "10 per cent" of progressive teachers who teach an original type, of course, but it is pretty certain to be relatively fair to the classes of the remaining 90 per cent of teachers.

It should be seen that textbook analysis reduces fundamentally to the method of pooled judgments of competent

persons, since each textbook's content represents the judgment of one or more supposedly competent persons.

An illustration of the textbook analysis method is given by the following quotation from the description of the validation of a general science test:¹

The selection of the items in the Ruch-Popenoe General Science Test has been based largely upon an analysis of the contents of existing textbooks in general science. Although there are more desirable criteria for the validation of test materials for most school subjects than the one here used, it is nevertheless true that, during the formative period of any new subject, teaching practice follows very closely the organization of the available textbooks. This very practical consideration is held to be sufficient justification for the construction of a test which embodies the best of the materials actually occurring in the books which must be used in general science classes.

The present form of the test is a revision and extension of the Range of Information Test in General Science published in 1919 by one of the authors. The original test consisted of three forms of fifty scientific terms each, which had been selected from a list of more than five hundred of such terms obtained by analysis of twenty-three texts and manuals in general science. Of these five hundred items, about 180 were common to half or more of the texts and for this reason were considered to be valid for test purposes. Some further eliminations, necessitated by reasons stated in the articles cited below, reduced the final number to 150. These were roughly equated and standardized as three forms of fifty items each.

In the present test, two of these forms have been standardized in more objective form. The ease of scoring has been greatly increased, and greater objectivity of scoring has been made possible.

The selection of the diagrams of Part II was carried out in a manner similar to the information item of Part I, except that the ratings of a number of experienced teachers were secured as an additional guarantee of the desirability and validity of these materials. Some further eliminations were made in order to prevent undue overlapping of the fields covered by Parts I and II.

¹ Manual of Directions for Ruch-Popenoe General Science Test (World Book Company), pages 1-2.

As a general summary of the textbook analysis method of validation, the following statements will probably suffice:

- (1) This method is not to be advocated where better methods can be applied.
- (2) In the case of subjects yet in the "embryo" stage, the method may be about the only available one.
- (3) The objections to be raised against the method are probably more serious in the eyes of the careful student of measurement than in the opinions of actual teachers, since tests so constructed do resemble the daily teaching materials.
- (4) Such tests are not likely to suggest teaching reforms to the same degree as tests that have been validated by the principle of social utility.

2. Analysis of Courses of Study

Validation by analysis of courses of study. Little need be said concerning this method, since it is in all important respects a variate of the textbook method. Its advantages and limitations are practically identical with those mentioned in the preceding section. On the whole, the analysis of courses of study is inferior in practice to the textbook analysis, due to the fact that courses of study in their usable (published) form are far less detailed than textbooks. Moreover, printed courses of study tend to deal with general aims, advice, and principles, and thus lack the concreteness necessary for the test builder.

3. Analysis of Examination Questions

Analysis of final examination questions. This method also is to be considered a variate of the preceding two. It has about the same limitations. Examination questions in comparison with textbooks as a source of valid test items

have one point of relative superiority. Textbooks contain many unimportant details over which mastery cannot be hoped for or even attempted. The teacher's selections of questions for her final examination represent an additional, thoughtful "culling out" of the non-essential. The teacher limited to five, ten, or twenty questions is likely to exert every effort to select the most valid content of the course. This selection tends toward added refinement over the textbook or course-of-study analysis method. It shades over at the same time to the method of judgments of competent persons.

The following quotation¹ presents an application of the method of analysis of final examination questions to the building of a test in high school biology:

The selection and validation of the items in the Ruch-Cossmann Biology Test have been largely based on the analysis and condensation of two thousand final examination questions obtained from a collection of final examinations used in all parts of the United States over a period of several years. This task, which alone occupied a year's time, has been summarized in an article by the writers in the *Journal of Educational Psychology* for May, 1924, entitled "Standardized Content in High School Biology." The following description of the method of determining the most valid materials for inclusion in a test of high school biology is quoted from the paper mentioned:

The final examination represents a more or less thoughtful attempt on the part of the teacher to isolate those facts and principles which can with most reason be expected to be retained by the pupil for some appreciable length of time after the completion of the formal work of the classroom and laboratory. It is for this reason that the authors chose to attempt to survey good teaching practice in biology by statistical analysis of final examinations collected from all parts of the country. The limitations of the method are as apparent as are its advantages, and no claim is made that finality of content determinations can or should be based upon this method.

¹ Manual of Directions for Ruch-Cossmann Biology Test (World Book Company), page 2.

Believing that the results of such a provisional investigation would possess enough merit to justify the effort expended, one of the authors (L. H. C.) began in 1918 actual work on the problem, the completion of which has for many reasons been sadly delayed. The method of attack will be very briefly presented in summarized form.

(1) A letter was sent to the Superintendent of Public Instruction in every state of the Union having a central officer, asking that he list the names of those schools and teachers who were doing the highest grade work in high school biology. Lists were received from 23 states, with a total of 126 schools answering to our specifications.

(2) Each of the 126 teachers was asked to send duplicate copies of all final examination questions used during the current year. Lists for previous years also were requested, and in some cases these were available for periods of five years or more prior to 1918. The total number of questions received ran slightly over two thousand.

(3) The two thousand questions were then analyzed for frequency and condensed and classified as 300 constantly recurring items. The group headings finally numbered eleven such phases of the total subject as: definitions of terms, identification of structures, life histories, diagrams, functions, economic and ecological facts, etc.

(4) The final list of 300 questions was then submitted to 100 of the leading teachers and authorities on the teaching of biology, for ratings as follows: (a) Rate "1" all questions which you consider to be entirely satisfactory and representative of the kind of knowledge which should be presented to high school students of biology, and which high school students can reasonably be expected to answer. (b) Rate "2" all questions which are partially satisfactory, but which you consider to be not highly desirable for any reason, . . . etc. (c) Rate "3" all questions which are entirely unsatisfactory.

Provision was also made for the writing in of any other questions not already listed. Sixty-eight teachers and 9 authorities (authors of texts, professors of the teaching of science, etc.) rated all the questions submitted. This was a remarkably high percentage of responses, and whatever selection did result from the failures to answer can safely be assumed to be a selection upward and hence could not have operated to decrease the reliability of the results.

The exact lists and ratings of most of the important items are reproduced in the article from which the above quotations are taken. For purposes of describing the validation of the Ruch-Cossmann Biology Test, it is sufficient to state that all test items are selected from classes "1" and "2" in the above rating scheme, the greater part being from class "1," plus a few additions from class "2" in the interests of increased difficulty. The test items therefore repre-

sent the combined best judgments of 77 competent teachers and authorities on the teaching of biology. This practically insures the worthwhileness of every test item included in the final forms of the test. Personal biases and hobbies were certain to have been ironed out by the mass judgments.

4. *Pooled Judgments*

Validation by pooled judgments of competent persons. Some of the uses of this method have been pointed out in connection with the preceding methods. In the last analysis, all validation methods reduce fundamentally to the consensus of competent opinion. The criteria of intelligence as a mental capacity represent merely a substantial agreement on the part of those psychologists who have given special study to that problem.

The curriculum, the course of study, and the textbook represent working agreements and the approximation to the best knowledge of the moment. This statement is introduced in support of the proposition that the educational method *par excellence* of today is fundamentally that of pooled judgments of educational authorities. This proposition is at the same time a justification of the use of judgments in test construction and a confession of the lack of a purely scientific technique for building either curricula or standard measurements in education.

The most obvious use of judgments in test construction arises in connection with the sifting of the original lot of tentative test items with a view to the elimination of the least valid materials. Experience has shown that the average or median judgment of a group of from three to ten careful judges is certain to be superior to the opinion of a single worker in approximating the true worth and difficulties of proposed test items. In the early stages of a test the rough work of eliminating unfit items is often accomplished by a

method of a double judgment upon (a) general merit and (b) difficulty. The most convenient way of doing this is to write each item on the face of a 3" \times 5" card, submitting the cards to each judge in turn with instructions somewhat as follows:

INSTRUCTIONS FOR RATING ITEMS

1. Study each item carefully and decide upon its general validity for inclusion in a test of , using a three-point scale as follows:

- 1 Entirely satisfactory
- 2 Fairly satisfactory
- 3 Objectionable; eliminate

2. Next examine the item again for its relative difficulty. Decide in which grade (age, etc.) the item is correctly located in the sense that approximately 50 per cent of pupils can answer the item correctly.

3. Turn over the card and record *both* ratings in the *upper right hand* (lower left hand, etc.; each judge being assigned a space for his ratings) thus:

$\frac{1}{VII}$ $\frac{3}{V}$ $\frac{2}{III}$, etc. Study the sample card below until you see how the recording is to be done.

The largest state in the Union is —		
California	New York	Texas
Montana	Pennsylvania	

(Front of card)

$\frac{1}{V}$	$\frac{2}{V}$	$\frac{1}{VI}$
$\frac{1}{VI}$	$\frac{1}{VII}$	$\frac{2}{VI}$

(Reverse of card)

4. Record nothing until you have made your decision about both ratings. You must take care not to turn over the card and expose the ratings previously given by other judges.

It is to be noted that each judge is assigned a special position on the reverse of the card for his judgments. In this way the source of the judgments can be traced. After

all ratings have been made, the median rating can be encircled with a red pencil as an aid in sorting the cards into piles of supposedly equal difficulty.

Ratings made in this way are certain to introduce a considerable amount of refinement in the preliminary sifting prior to actual try-out. Ratings pooled from three to ten judges often show correlations of from 0.60 to 0.80 or higher with the true order of difficulty as later found experimentally. The same is probably true for degrees of merit.

The scheme given above is not necessarily the best one possible, but it is simple and has repeatedly demonstrated its usefulness. The validity of the pooled judgments varies with the skill of the judges. It is probably true that the averaged judgments of six highly competent persons is worth more than those of a hundred non-expert persons.

The method of pooled judgments, alone or in combination with other methods, is by far the most common validation practice in educational test construction today. Unfortunately, the judgments very often represent the opinions of but one or two persons, and are not checked up against experimental findings. Study of dozens of educational tests published today has failed to find proof that any more rigid validation was undergone than that offered by the judgment of one or a very few judges.

5. *Rating Scales*

Rating scales in building up criteria. This method grows out of the one just described, from which it is distinguished only by virtue of certain special applications, chiefly in psychological test work.

In the construction of objective measures of psychological traits like will power, originality, leadership, etc., rating scales are often used. The method is so open to error that it is useful only in the hands of trained workers. The

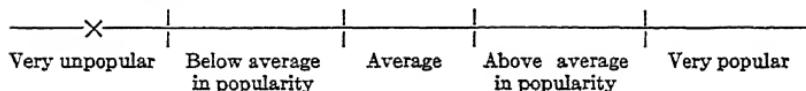
experimental studies of the validity of rating scales have been most discouraging. The widespread popular use of ratings can only be explained upon a basis of ignorance of the limitations of the method and the lack of better measures.

A method often used, and one which is superior to most other types of ratings, is illustrated by the following.

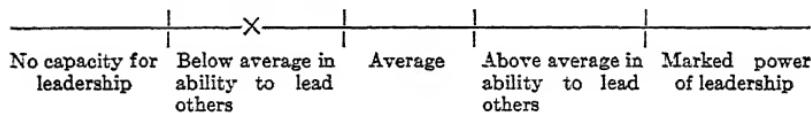
QUALITIES OF LEADERSHIP

DIRECTIONS. Place an \times on the line at the point which best describes the person rated. Do all the traits the same way.

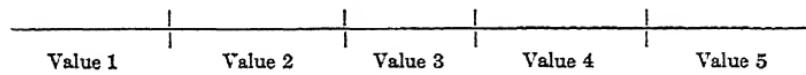
Trait 1. Popularity with other children



Trait 2. Ability to lead others



A number of such traits centering about the more general trait of "leadership" may be summed to form a single value. There are two ways of doing this: (a) by use of the values shown below, or (b) by measuring the distances of the \times from



the left end of the line and using this numerical statement of the distance as a rating.

Many other types of rating schemes have been used, some being variates of the method just described; but others are different in principle. Rating scales are highly *subjective* and hence open to errors of measurement to a marked degree (a phenomenon doubly subtle because of the apparent simplicity of ratings). Their use should be confined to a last resort.

Rules for securing ratings. A few rough rules may be laid down which will tend to minimize errors in the rating method:

- (1) Do not attempt to secure too fine gradations in judgment. A 5- to 7-point scale represents the practical limit of the human judgment to discriminate highly subjective facts. It might be noted that this is the reason why the letter-grade marking plan seldom considers more than five letter marks.
- (2) The validity of the rating varies with the *objectivity* of the trait rated. Traits like "physical beauty," "physical energy," although certain to show great variability in the ratings, will very likely tend to greater agreement than highly subjective traits like "sympathy" or "originality."
- (3) The validity of ratings increases with their *specificity*. Ratings on "love of nature" will be helped by obtaining separate estimates on five, ten, fifteen, or more specific facts like "interest in birds," "hiking in the woods," etc.
- (4) The combined judgments of a small number of intimate acquaintances are likely to prove superior to the averaged judgments of large numbers of casual acquaintances. This is equivalent to saying that the pooled opinions of three or four legal experts on an intricate point at law are better than those obtained by submitting the issue to popular vote.

Sources of error in rating scales. The pitfalls in rating schemes are in all probability numerous, but only a few have been studied experimentally. The following statements suggest some of the sources of error:

1. Judgments show *systematic biases* or *constant errors*; i.e., they tend uniformly to run too high or too low. A given

rater judges *all* the persons he rates too generously or too harshly, as the case may be. The best way to reduce errors of this sort is by pooling the judgments of many raters so that high raters and low raters tend to neutralize each other's errors. The following ratings illustrate the point.

TABLE 68

RATINGS OF THREE JUDGES ON TEN PUPILS FOR QUALITIES OF LEADERSHIP. (The scale used was: 10 = highest rating possible; 1 = lowest rating possible.)

PUPIL	JUDGE A	JUDGE B	JUDGE C	AVERAGE OF 3 JUDGMENTS
1. John Smith	10	2	8	6.7
2. Henry Jones	7	3	6	5.3
3. Frank Owens	6	2	4	4.0
4. Frederick Day	8	4	5	5.7
5. Oliver Holmes	6	2	3	3.7
6. William Weber	5	1	2	2.7
7. Kenneth Steele	9	7	5	7.0
8. Henry Forbes	10	6	10	8.7
9. Jesse Peters	4	1	1	2.0
10. Ellwood Brown	7	3	7	5.7
Average	7.2	3.1	5.1	
Average Deviation	1.6	1.5	2.1	
Standard Deviation	1.9	1.9	2.6	
Correlations:				
Judge A with Judge B	0.72			
Judge A with Judge C	0.88			
Judge B with Judge C	0.58			

The systematic tendency for Judge A to rate all subjects high is clearly evident. Judge B shows the reverse tendency. Since the correlations between ratings are moderately high, it is likely that the ratings are fairly accurate apart from these systematic errors.

2. Ratings show inequalities in the *degree of discrimination* used by different raters. Table 68 above shows this fact as

well if we examine the average and standard deviations. Judges A and B made their judgments pile up close to their averages. Judge C spread his ratings over a larger number of steps on the scale. Since the *effective* length of the scale is the number of steps actually used and not the number provided, considerable differences in discrimination are shown even with three judges. The probable reason for the lack of greater variability in most ratings is a psychological one; viz., the rater tends to avoid giving the extreme ratings on either end of the scale, not wishing to brand any one as "very high in leadership" or "very unattractive in appearance," etc.

3. Probably the most serious limitation of ratings lies in the phenomenon of "halo." By this term is meant the tendency to rate *all traits of the same individual* equally high or equally low — probably due to the influence of a general impression of, liking for, or attitude toward the one rated. The rater unconsciously seems to be "for" or "against" the person "on general principles."

A typical set of data showing marked halo effects is found in Boyce's study of teaching success by the score-card method.¹ A few of Boyce's results are worth quoting, because they touch upon several common educational practices (selection of teachers, grading pupils, etc., as well as test construction).

The suspicious thing about Boyce's ratings is the "dead level" uniformity. Everything is related to teaching success, and everything to pretty much the same degree. Undoubtedly halo will explain most of this situation.²

¹ Boyce, A. C., *Fourteenth Yearbook of the National Society for the Study of Education*, Part IV.

² See also Knight, F. B., "The Effect of the 'Acquaintance Factor' upon Personal Judgments." *Journal of Educational Psychology* (March, 1923), pages 129-142.

TABLE 69

Correlations obtained by Boyce between "general teaching ability" and:

1. General appearance	0.47
2. Health	0.56
3. Voice	0.53
4. Intellectual Capacity	0.62
5. Initiative and Self-Reliance	0.77
41. Attention and Response of Class	0.86
42. Growth of Pupils in Subject Matter	0.87
43. General Development of Pupils	0.88
44. Stimulation of Community	0.70
45. Moral Influence	0.71

Range: 0.38 (not shown in table) to 0.88

4. Another subtle and dangerous phenomenon of ratings arises from the fact that correspondence (high correlation) of the ratings of a large group of judges, even if very close, *does not prove that the judges are actually rating the trait which they are supposed to estimate*. The basis of the rating (and its reliability) may actually be some very different trait, but being common to all judges, it makes for high agreement (reliability). Thus, a group of raters might rate a class of pupils for intelligence (the pupils being total strangers, seen for the first time) and show considerable tendency to agree among themselves, without the ratings proving later to have any significant correlation with a good measure of intelligence (e.g., the Binet Scale). In such a case the agreement would necessarily be based upon the fact that the judges used the same, but spurious, criteria, such as personal appearance, vivacity, sparkling eyes, etc.

6. Correlations with School Marks

School marks as a criterion. The use of school marks in the validation of test materials is one of the most common

practices. Within the limits of the possibilities of the method, school marks form an excellent criterion.

It is obvious that a valid test in English or Latin should correlate fairly closely with the final grades earned in such courses. There can be little question that school marks are possessed of considerable validity. The same cannot usually be said for their reliability. The reliability of the end-term mark or grade in a school subject probably ranges between 0.50 and 0.75 as a rule. There is considerable experimental evidence on this point which suggests that these figures are conservative.

Returning to the possibility of marks as a criterion against which a test can be tried out, the usual procedure is that of correlating test scores and marks. Such correlations are never high, 0.85 being about the highest which the writers have ever seen reported in the literature for single classes. The reason for such low correlations with school marks is to be found in the low reliability of the marks. It can readily be shown that if school marks are no more reliable than 0.50 to 0.75, the highest possible correlation of a reliable test with such marks cannot exceed 0.70 to 0.85.¹

The use of school marks as a test criterion presents a simple and very generally useful method of validation in the case of educational tests within the limitations imposed by unreliability of the measures. This unreliability arises from several causes :

¹ It can be shown that the highest possible correlation which can be obtained with school marks, any test, or any other measure cannot exceed greatly the square root of the reliability coefficient of the measure in question. Thus if a given set of marks has a reliability of 0.64, the limit of correlation of *anything* with these marks is $\sqrt{.64} = 0.80$.

The above calculation also assumes the test to be perfectly reliable. This is never true. The limit of correlation of a test with 0.90 as its reliability coefficient, with school marks with a reliability of 0.64, is $\sqrt{.90 \times .64} = 0.76$.

For proof of these formulas, see Kelley, T. L., *Statistical Method* (The Macmillan Company, 1923), pages 205-208.

1. School marks are based upon tests, recitations, and examinations which are always *samplings* — never complete measurement in the sense that everything taught is fully measured. Marks cannot be perfectly reliable because they are samples, if for no other reasons.
2. Teachers' marks are open to the systematic errors of all ratings.
3. School marks are composite measures; i.e., they reflect behavior, work habits, mental ability, and other traits, as well as pure achievement.

The use of marks in the validation of mental tests is omitted from discussion here because of the intricacies of the arguments involved.

7. Rise in Percentage of Successes

Validation through percentage increases in successes. This provides one of the most effective methods of validating test items. It is limited in its utility only in two directions; viz., (a) in a few physiological capacities which do not continue to develop over a period of years, and (b) in those school subjects which are discontinuous over a series of grades.

The typical procedure is to give the experimental edition of the test to considerable numbers of pupils over a wide range of grades and ages. The pupils are then grouped by ages (or grades), and the percentage of pupils passing each test item at each age level is computed, with results like those shown in Table 70 and Figure 2 on page 321.

The graph allows the test maker to see at a glance certain facts about the difficulty and reliability of each test item.

Item 1 is characterized by a gradual but uniform rise in successes from Grades IV to X. It does not function at all in Grade III, and presumably not above Grade X. It is *normal* to Grade VII; i.e., about 50 per cent of successes. The

discriminative power of this item is not great; i.e., the slope of its curve is gradual. It will be a useful item because of the regularity and long-continuing rise.

Item 2 is harder than Item 1 and does not function until about Grade VI. It discriminates sharply in Grades VI to VIII but does not function below or above these limits.

Item 3 presents a common occurrence in test construction in that its grade rise is erratic. It rises and then falls, causing Grades IV, VIII, and X to show about the same per cent of successes. Such an item must be thrown away. The reasons for such "throwbacks" are not always evident. Sometimes, of course, they may be merely fluctuations due to small samples at each level. If 50 to 100 pupils are used at each grade level, marked irregularities must be attributed to faults within the item itself. In subjects like American history and grammar, where the teaching occurs only in occasional grades, these erratic effects may mean nothing more than alternating reviews and forgetting.

Item 4 shows no additional features worthy of comment. It is valueless below Grade VII, but functions nicely in Grades VII to IX.

Item 5 is a good item for the lowest grades, and probably in Grade II, which is not shown. It appears to be normal to the high third grade.

Two general procedures are open to the test maker in choosing items by this criterion, depending upon the range of grades for which the test is being constructed. If the range of grades is to be narrow (say III to V, or VI to VIII), the best thing to do is to select items like Nos. 4 and 5 which discriminate sharply. This will guarantee high reliability to the test when completed. If the test must be used over a wide range of grades, items like No. 1 will be better. The discriminative power of such items is not very great, due to the slow rise, but such items will function over the entire

TABLE 70

THE PERCENTAGE OF PUPILS PASSING TEST ITEMS AT EACH
SUCCESSIVE GRADE LEVEL

ITEM	PER CENTS PASSING								
	GRADE III	IV	V	VI	VII	VIII	IX	X	
1	0	1	15	28	48	67	86	98	
2	0	1	1	26	77	100	100	100	
3	0	10	38	58	40	11	24	12	
4	0	0	0	1	20	80	98	100	
5	20	68	95	99	100	100	100	100	

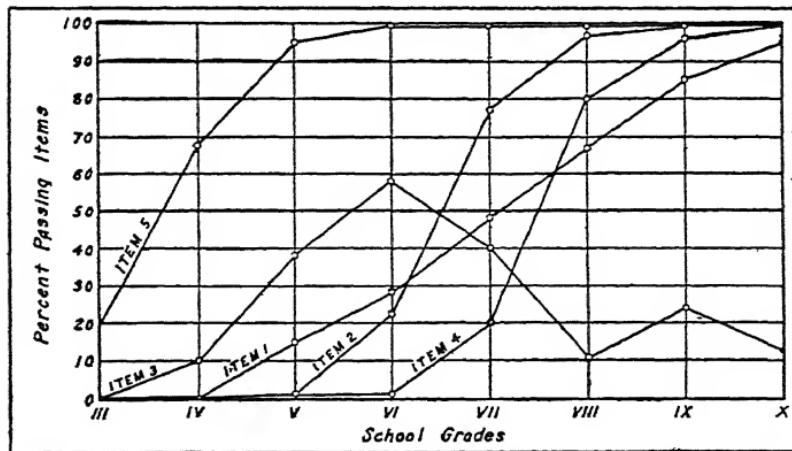


FIG. 2. Graphic representation of the facts of Table 70.

grade range. Such a test, when completed, will not show as high a reliability (per constant number of test items) as the former, but it will prove satisfactory.

The best practice for tests which must be employed over a wide grade or age range seems to be the use of a series of over-

lapping items, each with high slope. Figure 3 below shows how this is possible.

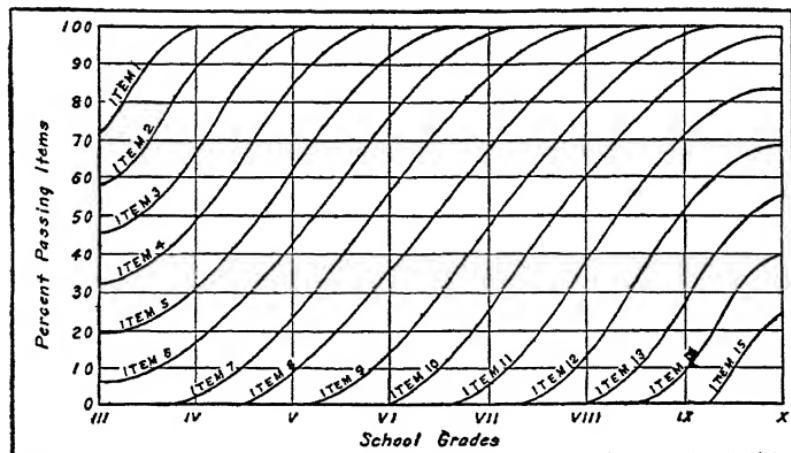


FIG. 3. Showing the placement of items to guarantee continued high discrimination and reliability.

If the items increase in difficulty uniformly and show sharp rises with each successive grade, the test must of necessity possess high reliability.

In summary, items with "throwbacks" are a source of great unreliability. Items passed by 0 per cent or 100 per cent are functionless but do not cause unreliability. They are to be looked upon as "dead timber." The sharper the rise, the greater the reliability of the item.

8. *Correlations against Validated Measures*

Validation against other validated criteria. The usefulness of this method is almost entirely confined to intelligence test construction, particularly group tests. Group tests are sometimes studied in the light of their correlations with the Binet Scale. A case in point, but here an individual test,

was Herring's Revision of the Binet Scale, the Stanford Revision forming the criterion.

9. The Method of Widely Spaced Groups

Validation of trade tests. Trade tests and many psychological tests have been validated by a principle termed "the method of widely spaced groups." During the World War a branch of the service was devoted to trade testing. Following the custom of labor organizations, men were classified upon the basis of training and experience as *experts*, *journeymen*, *apprentices*, and *novices*. These groups are roughly distinguishable, although much overlapping of abilities is the rule. On the whole, the averages of abilities of these four groups are certain to show a progression from novices to experts. In producing trade tests, large numbers of men were tested from each group, and tests showing the expected rise in average scores were held to be valid. In reality, *only four points* on a hypothetical scale were validated; the intermediate steps could be *assumed* to be valid. Figure 4 gives some notion of the significance of this procedure.

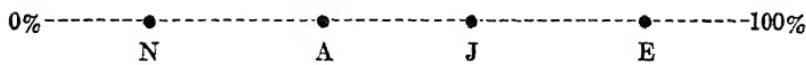


FIG. 4.

Points N, A, J, and E are the average scores of novices, apprentices, journeymen, and experts, respectively. Such points were experimentally determined. The dotted lines intermediate between these points are assumed to form a continuous scale.

A similar method has been used by Cady, Raubenheimer, and Cushing and Ruch in the attempt to validate tests of character traits.

The following diagram taken from Chapman¹ shows the differences in scores earned by a group of 80 men, a sampling of 20 each of novices, apprentices, journeymen, and experts.

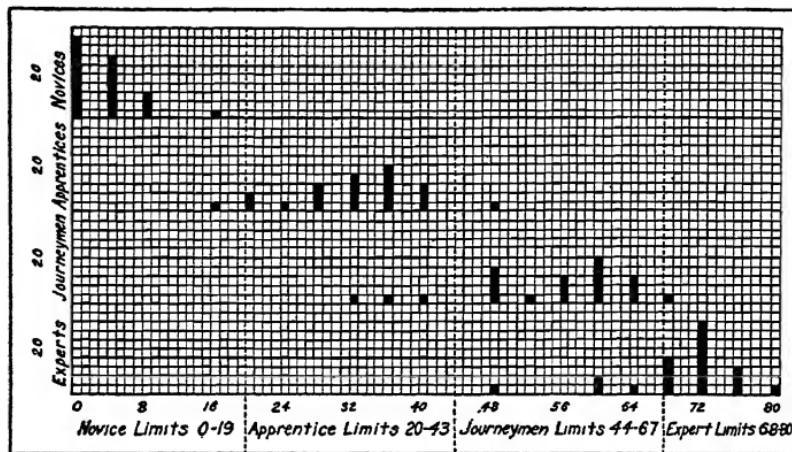


FIG. 5. Distribution according to individual total scores of 80 subjects used in standardization. Each square represents one man. The limits shown are the ones finally used in army testing.

Widely spaced groups in educational test construction. A useful variate of this principle allows its application to the problem of selecting educational test items of high discriminative value. A grade or each group can be subdivided into a "good" and a "poor" group by means of their total scores on the try-out form of a test in process of construction. Two test items have been selected for illustration. The experimental data follow:

PER CENTS PASSING

Grade.....	VII	VIII	IX	X	XI	XII
Item 1	27	39	52	67	89	98
Item 2	25	41	54	65	88	99

¹ Chapman, J. C., *Trade Tests* (Henry Holt & Co., 1921), page 106.

These two items appear to be about equally difficult and equally discriminating. If, on a basis of total scores on the test, the pupils, grade by grade, are divided into those above the median (the "good" group) and those below the median (the "poor" group), and the per cents are retabulated by these sub-groups, we find:

Grade.....	PER CENTS PASSING					
	VII	VIII	IX	X	XI	XII
Item 1 { "Good"	14	21	28	34	45	50
	"Poor"	13	18	24	33	44
Item 2 { "Good"	16	26	31	37	50	61
	"Poor"	9	15	23	28	38

It is evident that Item 2 has much more value in discriminating small differences in abilities at a given grade than has Item 1. Approximately equal in other respects, Item 2 is greatly to be preferred over Item 1.

10. Validation by the Principle of Social Utility

The criterion of social utility. The determination of the social usefulness of educational content is primarily a problem in curriculum construction. This fact probably explains the reason that tests have not usually been checked in the light of the social significance of the information called for by the test. In a few subjects, chiefly elementary, important experimentation has been done in this direction. Extensive counts have been made by Thorndike and Horn on vocabularies; by Horn and Ashbaugh on spellings used in business; and by Wilson and others on the arithmetic of business. A number of major investigations in Latin, modern languages, and English are under way. The vocabulary count of Thorndike is of incalculable aid in devising reading tests, and the recently completed larger study by Horn will be of even more service. A number of new tests will be standardized in the course of the investigations of the

Modern Foreign Language Study. Some of these will be based upon determinations of social utility.

11. *Psychological and Logical Analysis*

Validation by psychological analysis. A few psychological tests rest their claims to validity upon facts established in the psychological laboratory. A typical example is found in the Seashore Measures of Musical Talent. Seashore essays the following analysis of the musical mind:¹

FACTORS OF THE MUSICAL MIND

I. Musical sensitivity

- A.* Simple forms of impression
 - 1. Sense of pitch
 - 2. Sense of intensity
 - 3. Sense of time
 - 4. Sense of extensity

- B.* Complex forms of appreciation
 - 1. Sense of rhythm
 - 2. Sense of timbre
 - 3. Sense of consonance
 - 4. Sense of volume

II. Musical action

Natural capacity for skill in accurate and musically expressive production of tones (vocal, instrumental, or both) in:

- 1. Control of pitch
- 2. Control of intensity
- 3. Control of time
- 4. Control of rhythm
- 5. Control of timbre
- 6. Control of volume

¹ Seashore, C. E., *The Psychology of Musical Talent* (Silver, Burdett & Co., 1920), pages 8-9.

III. Musical memory and imagination

1. Auditory imagery
2. Motor imagery
3. Creative imagination
4. Memory span
5. Learning power

IV. Musical intellect

1. Musical free association
2. Musical power of reflection
3. General intelligence

V. Musical feeling

1. Musical taste
2. Emotional reaction to music
3. Emotional self-expression in music

Based upon this analysis into underlying musical capacities, Seashore has developed his Measures of Musical Talent described elsewhere in this volume (see page 192).

12. Validation by Correlation with Other Measures

Test intercorrelations as criteria. In developing certain new types of tests of a psychological character, it is often necessary to establish a unique type of validity which may be called its "non-duplicative nature." The human mind seems to be a complex of general and special abilities. In working out new tests in the field of character and personality traits, non-intellectual traits, mechanical abilities, musical and artistic abilities, etc., it is valuable to correlate the scores of the new test with a wide variety of previously existing tests, scales, and measures. This rather random procedure often serves to establish the fact that the new measure is not a mere duplication of older tests but that it

“taps” a new field of the human mind. This intricate and controversial field is outside the scope of this volume and should be reserved to the experimentalist and statistician.

Chapter XVIII continues the discussion of the construction of educational and mental tests, with special reference to the experimental try-out of the items surviving the preliminary validation described in the present chapter.

CHAPTER EIGHTEEN

THE CONSTRUCTION OF EDUCATIONAL AND MENTAL TESTS (*Continued*)

THE EXPERIMENTAL TRY-OUT OF ITEMS

The preliminary try-out. The preceding chapter incidentally has suggested the general procedure in handling the first experimental edition of the test. As soon as a sufficient number of items have survived the rating process, these are assembled and printed as the first trial edition. The actual numbers of items to be included in the first experimentation should be *at least twice* the total number contemplated for all final forms. At times, three or four times the projected final numbers are tried out. This surplus allows the elimination of all items which show the faults of erratic rise, slow rise, too great difficulty, too great ease, etc.

From 100 to 200 pupils per grade will ordinarily be a sufficient sampling for this first try-out. The selection of such pupils need not be as carefully controlled as will be necessary in the later try-outs, but it should be done with reasonable care.

The following statements will be suggestive of the typical procedure:

1. The pupils should be drawn from as typical and normal school conditions as possible. It is better to use *all* pupils in each grade, since to entrust to the teacher the selection and reduction of pupils to the desired numbers is likely to result in a more or less selected and unrepresentative group. If only a portion of the pupils in each grade can be used, one of the safest procedures is: (a) Have the teacher rank the pupils in order of their accomplishment, and then (b) start with the median or middle pupil and work equally far both

ways until the required numbers are selected. This plan will tend to insure that an average group is finally obtained. Another possibility, better in some ways, is to take every second, third, fourth, etc., pupil after the ranking has been done by the teacher. A sampling by alphabetical order can be handled in a similar fashion.

2. On the first try-out allow enough time for every pupil to attempt each item. This will require considerably more time than is planned for the final time limits. If conditions will not permit so much time, a fair compromise would be to stop the examination when 90 per cent of the pupils have attempted all the items. A large amount of time is needed for the first try-out, because items near the end of the test will otherwise not be attempted by many pupils and will thus tend to decrease the per cents of pupils passing on these items. The end items will appear to be much more difficult than is actually the truth.

3. It has already been suggested that the items in this try-out should be arranged in approximate order of difficulty by means of pooled judgments. The placing of the easy items first and the difficult items last accomplishes two desirable things: (a) the prevention of discouragement early in the test and (b) economy of time in answering.

4. If rigid economy of time and expense demands an abbreviated procedure, the first try-out may be confined to alternate grades; e.g., VII, IX, and XI, or VI, VIII, X, and XII.

5. If the preliminary test has been broken into two or more forms for the sake of ease of arranging the test sittings, the order of giving these forms is important in order to equalize practice effects. If we assume that there are three preliminary forms to be tried out, the following is a satisfactory plan:

$\frac{1}{3}$ of each grade should take the forms in the order A-B-C

$\frac{1}{3}$ of each grade should take the forms in the order B-C-A

$\frac{1}{3}$ of each grade should take the forms in the order C-A-B

(The capital letters designate the three forms.) Practice effects will be neutralized by this procedure for all practical purposes.

Computation of Item Difficulties

Determination of per cents passing each item. After the preliminary form or forms have been given and scored, the per cents of pupils passing each item grade by grade (or age by age) are computed and tabulated. Chapter XVII has explained the use of such tabulations in the elimination of the unfit items. This culling out is continued until the requisite final numbers remain; i.e., exactly enough for all final forms. The final total number is determined variously, the following being the guiding principles:

1. *Desired reliability.* The careful test maker sets a minimum standard of reliability to be reached. Through correlations of the preliminary forms, it can be estimated about how many items will be needed to give a reliability of 0.75, 0.80, 0.90, as the case may be, for the grade or age groups represented. A margin of 10 per cent to 20 per cent of extra items should be allowed for safety.

2. *Printing requirements.* It is necessary at this stage that consideration be given to the space requirements of the tests in final printed form. This factor aids in deciding the numbers of items to be retained.

The per cents passing each item, grade by grade, are tabulated in a form like the one shown on the following page, and in addition, each item is plotted on a graph in careful experimentation. The graphic presentation brings out no facts not covered by the tabulation, but it is easier thus to visualize the behavior of the items.

TABLE 71

PER CENTS OF CORRECT RESPONSES ON PRELIMINARY FORM

Grade	VII	VIII	IX	X	XI	XII
Item 1	52	69	81	95	99	99
Item 2	93	99	100	100	100	100
Item 3	16	29	41	57	77	91
Item 4	99	99	100	99	98	100
Item 5	40	53	70	89	99	100
Etc.						

(See also Table 70 and Figures 2 and 3 in Chapter XVII.)

The unsatisfactory items are then crossed out of the tabulation, leaving the final number agreed upon. If the test items are not to be scaled individually (i.e., given weights for importance), the items are now ready for breaking into the first set of approximately equivalent forms.

Scaling or Weighting of Items

Methods of weighting items. Tests and scales may be classified with respect to weighting as of three types:

1. Quality scales with weighted values
2. Product scales with weighted values
3. Product scales with unweighted values

Each of these will be treated briefly in turn.

Quality scales. The best examples of quality scales are to be found in existing handwriting and English composition scales. Each item (composition or specimen of handwriting) is given a weight upon the basis of ascertained merit. The method of weighting is to be regarded as a statistical refinement of the method of pooled judgments. The items are usually submitted to 100 or more competent persons who assign ratings to each specimen. The ratings are then

pooled in a special manner and the weights determined from appropriate tables.

Omitting many details, the general procedure is:

Step 1. Instruct the judges to rate each specimen upon a scale of 1, 2, 3, . . . 10 points for general merit.

Step 2. Tabulate the ratings for all specimens and all raters.

Step 3. Take each specimen in turn and determine the per cent of judgments by which that specimen is better than each of the others.

Step 4. Rearrange the specimens in order of merit, placing the per cents in a tabulation similar to the following:

53%	of the judges agree that Specimen J is better than Specimen D
58%	of the judges agree that Specimen B is better than Specimen D
65%	of the judges agree that Specimen F is better than Specimen D
69%	of the judges agree that Specimen A is better than Specimen D
78%	of the judges agree that Specimen C is better than Specimen D
83%	of the judges agree that Specimen G is better than Specimen D
94%	of the judges agree that Specimen E is better than Specimen D
97%	of the judges agree that Specimen I is better than Specimen D
100%	of the judges agree that Specimen H is better than Specimen D

Step 5. Assign the weights corresponding to the per cents of "better than" judgments in the Fullerton-Cattell table¹ given below (Table 72).

It is not necessary for the reader to understand the derivation of this table in order to follow the general outline of the theory underlying its use. It will help in the discussion to change, at this time, the hypothetical results listed under Step 4 to the weights given by Table 72 on the following page.

¹ This table has been variously quoted and reproduced. As far as the authors can determine, it was originally published in the *University of Pennsylvania Philosophical Series*, No. 2 (1892), in an article by J. S. Fullerton and J. McKeen Cattell, on "The Perception of Small Differences."

A more convenient reference for this table is to be found in Thorndike, E. L., *Mental and Social Measurements*, page 123.

TABLE 72

A TABLE FOR ASSIGNING WEIGHTS TO TEST ITEMS UPON THE BASIS
OF THE PERCENTAGES OF JUDGMENTS THAT $x > y$

% r	$\frac{\Delta}{P.E.}$						
50	.00	60	.38	70	.78	80	1.25
51	.04	61	.41	71	.82	81	1.30
52	.07	62	.45	72	.86	82	1.36
53	.11	63	.49	73	.91	83	1.41
54	.15	64	.53	74	.95	84	1.47
55	.19	65	.57	75	1.00	85	1.54
56	.22	66	.61	76	1.05	86	1.60
57	.26	67	.65	77	1.10	87	1.67
58	.30	68	.69	78	1.14	88	1.74
59	.34	69	.74	79	1.20	89	1.82
						90	1.90
						91	1.99
						92	2.08
						93	2.19
						94	2.31
						95	2.44
						96	2.60
						97	2.79
						98	3.05
						99	3.45

Looking up each in order, we have:

Specimen D .00 (arbitrarily given)

Specimen J .11

Specimen B .30

Specimen F .57

Specimen A .74

Specimen C 1.14

Specimen G 1.41

Specimen E 2.31

Specimen I 2.79

Specimen H ∞ (Indeterminate; might be set arbitrarily to 4.0 or some other value.)

The weights are the values in the $\frac{\Delta}{P.E.}$ column opposite the obtained percentage of judgments by which each specimen was better than Specimen D.

The theory underlying this method is that when 50 per cent judge $x > y$, there is no ascertainable difference in merit between the two. As the percentage rises above 50 per cent, we can be more and more certain that $x > y$. Finally, when 100 per cent agree that $x > y$, the chances

become infinitely great that x is really superior to y . As seems logical, the weights become increasingly larger in their increments as the $\% r$ approaches 100 per cent.

This method has been used with modifications in a variety of tests and scales, the Lewis and the Hillegas composition scales being well-known examples.¹

Product scales with weighted values A number of common high school tests of the familiar type have weighted values for assigning the credit earned by passing each item. The Henmon French and Latin tests, the original edition of the Douglass Algebra Test, the Iowa Physics Scales, and many others allow extra credit for passing the difficult items by means of a system of weights. Theoretically, it is sound to give extra weight to difficult items provided (a) the differences in difficulty can be fairly evaluated in quantitative terms, and (b) all items can be assumed to be equally *valid* in all respects other than difficulty alone. Condition (a) can probably be met satisfactorily by the use of appropriate tables of the probability integral.² Condition (b) contains possibilities for error in weighting that statistical methods are powerless to prevent; viz., *the weights do not distinguish between intrinsic difficulty and the low percentage of successes arising from the relative rarity or unimportance of the information called for*. Thus a difficult item of great importance may be weighted about the same as an easy item of rare occurrence and hence unknown. This is equivalent to saying that low percentages of correct responses actually arise from various causes, but the method of weighting treats all items as if the only cause of inequality in percentages of responses was that of intrinsic difficulty.

¹ Lewis, E. E., *Scales for Measuring Special Types of English Composition* (World Book Company, 1921), pages 11-26.

² Rugg, H. O., *Statistical Methods Applied to Education* (Houghton Mifflin Company, 1917), Table VI, pages 396-400.

Consider the three following arithmetic problems:

1. A man had 160 acres of land and then inherited 80 acres more from his father. How much land did he then own?
2. A rectangle 3" by 4" was divided into two right triangles by means of a diagonal. What is the length of the diagonal?
3. The cubes of two numbers stand in the ratio of 512:3375. What is the ratio of the two numbers?

Suppose that these three problems had been given to 1000 eighth-grade pupils, with the following results:

	WEIGHT
Problem 1. 99% passed; 1% failed72
Problem 2. 85% passed; 15% failed	1.97
Problem 3. 1% passed; 99% failed	5.28

The weights are read from Table VI in Rugg. The third problem would receive a larger weight than Problem 2 — probably because it is never taught. Here, unimportance is given extra credit.

It is significant that Douglass,¹ after using weighted values with the first edition of his algebra tests, discarded the weights in his revision and extension of his tests. He found the correlation of weighted and unweighted scores ranged from 0.91 to 0.99 and concluded that no real advantage attached to the weights commensurate with the extra labor which the weights involved in computing the scores.

On the whole, the best modern practice is away from the use of weights in ordinary product tests.

Unweighted product scales. It is a well-known fact that the scale of abilities represented by the raw point-scores earned on a test is an arbitrary one. The test units do not start from a true zero-point, and the differences in educational

¹ Douglass, H. R., and Spencer, P. L., "Is It Necessary to Weight Exercises in Standard Tests?" *Journal of Educational Psychology*, Vol. XIV (1923), pages 109-112.

accomplishment are not necessarily the same between the scores 10 and 20 and between 90 and 100. A change of score of 2 points in one part of the arbitrary scale may equal a change of 10 points in another part of the scale.

In the immediately preceding pages, two methods of equalizing such inequalities have been described. Most tests which contain a large number of items do not attempt to equalize such differences directly, because it is much simpler and probably quite as accurate to allow the norms to equalize the arbitrariness of the scale of raw scores. Educational measurement has never developed true units like those of the physical sciences. The nearest approach to a mental or educational unit is perhaps to be found in the concepts of *educational* and *mental age* increments. It is assumed that the development of a child, from one year to the next, is roughly equal, at least up to a certain point. This may or may not be true, but experience has shown it to possess sufficient truth to be a useful way of treating test scores. The advantage of age norms (grade norms are similar here) is that the unequal increments in the raw point-scores are referred to the age scale, and thus tend to have more equal units than if left in terms of the arbitrary units of the test.

By giving the test to large numbers of pupils at each age (or grade) and computing the average score per age, the raw scores become transmuted to age scores; the inequalities in the increments age by age or grade by grade thus tend to disappear from consideration.

Everything considered, the best present practice is to leave test items unweighted and to approximate equality of units by means of tables of norms. These may be grade, age, percentile, or other types of norms.

II. BREAKING THE TEST ITEMS INTO EQUIVALENT FORMS

Methods. The next step is that of breaking the items into equivalent forms, usually two in number. This is a very simple procedure, although one that must be carried out in a systematic manner if the two new forms are to be really equivalent. After all eliminations have been made, the items should be arranged in the order of increasing difficulty — i.e., decreasing percentages of successes. The items are then renumbered 1, 2, 3, 4, etc., in ascending order of difficulty. The two new forms, which we can designate as *A* and *B*, should be made up as follows:

FORM A	FORM B
1	2
4	3
5	6
8	7
9	10
12	11
Etc.	

This plan should be examined carefully in order that the significance of the method may be seen. To throw the odd-numbered items into one form and the even-numbered items into the other form would have the result of making the odd form markedly easier than the even form, due to the fact that each odd item would be slightly easier than its paired even item. The above scheme places exactly the same number of odds and evens in each form.

The means and standard deviations of the new forms should prove to be almost exactly equal. If not, they can be adjusted in the second try-out. Chapter IV has already presented the definition of equivalence of forms and an outline of the conditions which must be met.

The Second Try-out: Equivalence of Forms

Testing equivalence. The two new forms must be given an experimental try-out to make certain that the forms are equivalent. The numbers used in this stage should be larger than for the original try-out of items. All grades or ages should be included.

If the forms prove to be sufficiently equal in means and standard deviations to stand in their present form, the data from this try-out can be used in helping to establish norms. Even if slight readjustments are needed and a few items have to be shifted from one form into another, the results from this try-out may be used in building up norms.

The means and standard deviations should agree to a degree represented by a difference not exceeding 0.5 point at any age or grade level if the same set of norms is to be used for both forms.

In case there is marked lack of equivalence of forms, it will be necessary to redetermine the percentage of successes at each level exactly as in the first try-out.

Determination of Final Time Limits

Experimental methods. The final setting of time limits must be done experimentally. Ordinarily it can be done in connection with the second try-out by the use of certain devices. Before these are described, the principles involved in fixing time limits should be stated.

There are two roughly distinguishable types of tests, *speed* tests and *power* tests. The former have sharp time limits, and the scores are intended to show *quantity of accomplishment per unit of time*. Only the most rapid pupils, or none at all, are expected to finish. These speed scores are probably useful in a few subjects like reading and arithmetic, but they are of doubtful validity in most high school subjects. Speed,

per se, except in occasional life occupations, does not seem to have high social utility. On the whole, the emphasis on speed should be thought of as evidence of mastery rather than as an end in itself. Speed scores are also usually less reliable than accuracy scores in most test situations. Power tests are intended to reveal, not how fast a pupil can work, but *what is the most difficult task which he can perform*. For purposes of power testing, time limits should be very generous. Theoretically, there should be no time limits at all. In actual practice it is necessary to set some maximum working time, since there are always a few pupils who will keep on working indefinitely, and long after they have ceased to accomplish anything. A practical solution of the matter of timing power tests is to set the time limits at a point which will allow 90 per cent to 95 per cent of all pupils to attempt each item.

The determination of such a limit can often be managed by the use of two or three different colors of lead pencils. The first try-out should have given a rough notion of the length of time needed to answer a given number of items. Suppose that such knowledge led to an estimate of 25 minutes as the possible time needed for 90 per cent to finish. The pupils could be given three pencils when the blanks are passed out, a black-, a red-, and a blue-leaded one. They are allowed to work for 20 minutes with black pencils. The signal is then given to change to the red. After 5 minutes more, they are directed to change to the blue pencils. Subsequent examination of the blanks will show about what the time limits should be. If desired, the scores may be computed on a 20-, 25-, or 30-minute basis. This method often permits this try-out to aid in establishing norms.

A second and more exact method (but one which does not allow the use of results for norms) is that of writing the elapsed time, every half minute, upon the blackboard. As

each pupil finishes the test he writes the figure then appearing on the blackboard upon his paper. By making a frequency distribution of these working times, any desired degree of accuracy can be had for timing the tests in final form.

Time limits as a source of error in test scores. There are many test critics who hold that time limits are a source of great injustice to the "slow but accurate" pupil. This is probably true in some cases, how true depending upon the sharpness of the time limits. The 90 per cent to 95 per cent rule just given will eliminate most of the force of this criticism. Some writers, chiefly laymen in the test field like Walter Lippmann, have claimed that even the poorest student, if given sufficient time, would come to make a high score on many speed tests. Lippmann had in mind the army intelligence examination Alpha, but other writers have extended his criticisms to all speed tests. For this reason it may be justifiable to review the experimental evidence on the point. This evidence, although not extensive, points uniformly to the same conclusion.

"Speed" vs. "power" in tests. May and Terman found a correlation of 0.965 upon 510 soldiers for Army Alpha given with regular time limits (which are pretty sharp) and with double time allowances.¹ Ruch and Koerth repeated the same experiment, using 122 college freshmen, keeping separate scores for regular time, double time, and unlimited time. The same test was used. The results were:

	P.E.
Single time with double time	$0.966 \pm .004$
Single time with unlimited time	$0.945 \pm .007$

Some of the subjects worked four or five times the regular time, but in no case did any of the poorer ones (under regular time limits) even approach the scores of the better ones.

¹ "Psychological Examining in the United States Army." *Memoirs of the National Academy of Science*, Vol. XV (1921), page 416.

The gains among those who made high scores under regular conditions were nearly as great under unlimited time as among those who started low, in spite of the fact that the superior students were so near the top of the scale that they had little opportunity to earn additional points.¹

Ruch has also experimented with the Terman Group Test of Mental Ability and the Stanford Achievement Test in the same way, with the following results:²

CORRELATION OF REGULAR TIME AND UNLIMITED TIME SCORES ON
MENTAL AND EDUCATIONAL TESTS

TEST	<i>r</i>	<i>N</i>
Terman Group Test	$0.960 \pm .004$	150
Stanford Reading Examination .	$0.968 \pm .005$	64
Stanford Arithmetic Examination	$0.976 \pm .004$	86

In general, then, the evidence shows there is little danger of great injustices being done by the use of time limits, provided these are reasonably generous.

Chapter XIX continues the discussion of the construction of educational and mental tests, with special reference to the topics of the derivation of norms and the determination of reliability.

¹ *Journal of Educational Psychology* (April, 1923), pages 193-208.

² *Journal of Educational Research* (January, 1924), pages 39-45.

CHAPTER NINETEEN

THE CONSTRUCTION OF EDUCATIONAL AND MENTAL TESTS (Continued)

III. THE DERIVATION OF NORMS

The final try-out of the test. After the equivalence of the forms of the test has been established and equitable time limits determined, the test forms are ready for the final derivation of norms or standards. It is assumed that the final grade and age range to be covered by the test has been fixed by this stage. There remains only the task of deciding upon the kinds of norms to be assembled and the numbers of cases to be used. Chapter IV in Part I has commented at length upon the need for *representative* sampling in contrast with the practice of placing faith in large numbers alone.

It is best to give the first form of the test to half of the pupils, following with the second form, and to reverse this order with the other half of the pupils. The halving should, of course, be done within each age or grade group. If age norms are contemplated, all pupils within the age limits covered by the test should be tested regardless of the grade in which they are located. The practice of giving the forms in alternate orders to halves of the total group is designed to neutralize possible practice effects.

The further matters of norm derivation will be brought out in connection with the discussions of the various possible types of norms.

Types of norms. Four principal kinds of norms will be discussed :

1. Grade norms
2. Age norms
3. Percentiles
4. *T*-scores and other standard measures

1. *Grade norms.* These have been by far the most widely used type of norm in the past, and will probably always be demanded for elementary school tests. Age norms are coming to supplement grade norms and probably will finally be generally preferred. Grade norms have relatively less utility in high school subjects than in the lower grades.

Grade norms are easy to interpret and are very simple in their calculation, being either averages (means) or medians of unselected grade groups. As has been mentioned repeatedly, the only difficulty in obtaining grade standards arises from the task of selecting representative samplings of pupils at each grade. The dangers arising from undue faith in numbers alone, from "stacking" the results by over-emphasis on city school returns, from errors introduced by differences in grade classification, and from the fact that certain parts of the country have but seven elementary grades, have been mentioned. Aside from the added expense involved, it is far better to gather norms directly than to await returns sent in by users of the tests.

The main advantages of grade norms may be summarized once more, as follows:

- (1) They are easily derived, being simple averages;
- (2) The grade concept is a familiar one in the minds of school officers;
- (3) The grade is the unit of school classification;
- (4) Wide applicability.

The main limitations are:

- (1) Grade location is a man-made thing, and the average maturity of pupils, grade by grade, differs in different schools; i.e., the percentages of acceleration, at-gradeness, and retardation are not constant from school to school;

- (2) Grade standards have relatively little usefulness in the high school;
- (3) Grade norms do not offer easy comparisons with mental test results, since the latter are almost invariably expressed in age units;
- (4) Grade norms allow certain subtle errors in interpretation unless age and mentality are taken into account as well. (See Chapter II, Part I.)

A recent innovation in the use of grade norms has been inaugurated by McCall in the so-called *G*-score. The *G*-score is found by dividing the increment between grades into tenths. The pupil's point-score can then be translated into tenths of grades, like 4.2, 8.6, or 11.5, etc. This is a useful device and is to be regarded as a further development of the familiar practice of supplying norms by half-years. There is one possible danger in the *G*-score; viz., that the use of such minute differences as tenths of grades may tend to give a spurious sense of refinement in our thinking. It can be shown, as was evident from much of the data in Part II, that the probable error of a test score is very often a large fraction of the increment between two successive grades. When this is true, the refinement of the *G*-score loses all meaning for individual pupils. It may still be useful for class averages even in such cases.

2. *Age norms.* These are the arithmetic means (more rarely, medians) of as nearly unselected age groups as can be obtained. The usual procedure is to test every child in school, regardless of grade, whose age falls within the limits for which the test is planned. Moreover, it is desirable to extend the ages tested at least one, and preferably two, years above and below the limits actually settled upon.

The experimental facts needed for the calculation of age norms can be illustrated by the following data for the Stanford Achievement Test:

Age group	9	10	11	12	13	14
Mean score	23	33	46	57	66	72

Since the 9-year-old group will average 9-6 (9 years, 6 months) in age because it includes pupils 9, but not yet 10, years old, the raw score, 23, is the age equivalent of 9-6. The score, 33, is the age equivalent of 10-6.

The method of finding the intermediate months of EA (Educational Age) is that of interpolation between the experimentally determined values for successive ages, thus:

Raw score	23	24	25	26	27	28	29	30
	9-6	9-7	9-8	9-10	9-11	10-0	10-1	10-2
	31	32	33					
	10-4	10-5	10-6					

The values in bold-faced type are those found by experimentation; those in light-faced type were found by interpolation.

Fairly unselected age groups can be found in the public schools between the ages of 8 and 14; above and below these limits considerable portions of the groups are not in school. Twelve-year-olds form the largest age population of our schools and hence are probably the least highly selected group.

The chief advantages of age norms are:

- (1) Age is a natural unit not affected by school classifications;
- (2) Age equivalents permit easy comparisons with mental test results (mental ages);
- (3) Age standards tend to reveal faulty grade classification, effects of excessive retardation, etc.;
- (4) Age is probably more constant in meaning in all parts of the country;
- (5) Age norms allow comparisons with chronological ages — a necessary thing in interpreting test scores.

The main limitations of age scales are:

- (1) Greater experimental difficulty in obtaining norms;
- (2) Age is less familiar in our thinking about school achievement than grade equivalents;
- (3) Age norms have little usefulness in high school subjects.

3. *Percentiles.* Percentiles, while less well understood by teachers, pupils, or parents, have a great many advantages for use with high school tests. The statistical calculation of percentiles, although a simple matter to the statistician, is not sufficiently well known among educators to guarantee full understanding of the meaning of such values. For this reason Table 73 presents a tabulation of the scores of 1036 pupils on the Ruch-Popenoe General Science Test. The tests were all given at the end of nine months of study of the subject. Several dozens of separate schools and more than a dozen of the states of the Union are represented.

The distribution has been grouped in units of five. The highest class interval runs from 77.5 to 82.5, and includes all scores from 78 to 82, inclusive. The limits of the intervals are carried to five tenths of a point, upon the reasoning that the limit between two successive class intervals is exactly halfway between the lowest value falling in the one and the highest value falling in the next lower class.

The column headed *f* states the number of pupils earning scores falling within each class interval.

The column of *cumulative frequencies* *g* gives the total number of cases falling above each successive class limit. Inspection of the values in this column will show that they are running sums of the frequencies as we move down through the class intervals.

The calculation at the right shows the exact details of figuring percentiles. Every tenth percentile (i.e., *decile*) has been computed. To illustrate, the following description

TABLE 73

SHOWING THE CALCULATION OF PERCENTILES FOR 1036 SCORES ON
THE RUCH-POOPENOE GENERAL SCIENCE TEST

CLASS INTERVAL	<i>f</i>	CUMULATIVE FREQUENCY	CALCULATION OF PERCENTILES
77.5-82.5	1	1	90-percentile = $57.5 - \frac{(103.6-64)(5)}{52} = 53.7$
72.5-77.5	3	4	
67.5-72.5	11	15	80 " = $47.5 - \frac{(207.2-206)(5)}{127} = 47.5$
62.5-67.5	24	39	
57.5-62.5	25	64	70 " = $47.5 - \frac{(310.8-206)(5)}{127} = 43.4$
52.5-57.5	52	116	
47.5-52.5	90	206	60 " = $42.5 - \frac{(414.4-333)(5)}{129} = 39.3$
42.5-47.5	127	333	50 " = $37.5 - \frac{(518.0-462)(5)}{153} = 35.7$
37.5-42.5	129	462	
32.5-37.5	153	615	40 " = $32.5 - \frac{(621.6-615)(5)}{177} = 32.3$
27.5-32.5	177	792	
22.5-27.5	151	943	30 " = $32.5 - \frac{(725.2-615)(5)}{177} = 29.4$
17.5-22.5	66	1009	
12.5-17.5	25	1034	20 " = $27.5 - \frac{(828.8-792)(5)}{151} = 26.3$
7.5-12.5	2	1036	10 " = $27.5 - \frac{(932.4-792)(5)}{151} = 22.9$

is given for the values occurring in the calculation of the 90-percentile:

- 57.5 is the upper limit of the class interval containing the percentile sought (90-percentile)
- 103.6 is $1036 \div 10$, or the number of cases falling above the 90-percentile
- 64 is the cumulative frequency above the interval containing the 90-percentile
- 5 is the range of the class interval (the grouping)
- 52 is the frequency of the interval containing the 90-percentile

Translating this into words, we can reason as follows: We wish to find the 90-percentile; that is, a point in the distribution above which 10 per cent of the total cases fall and below which 90 per cent of the cases are located. One tenth of 1036 is 103.6. We must therefore move down the column of cumulative frequencies and decide in which class the 90-percentile falls. This we find to be the class, 52.5–57.5. Above this class there is a total frequency of 64. We must therefore move downward into this class interval 103.6 minus 64 frequencies. The next question is: What fraction of the distance across this interval (5 points) is 103.6 minus 64? The frequency within the interval is 52. The answer to our question is therefore:

$$\frac{(103.6 - 64)(5)}{52} = \frac{198}{52} = 3.8 \text{ units}$$

The factor 5 is introduced because the distance across the class interval is 5 points, due to the method of grouping used. The last step is to subtract the 3.8 from 57.5 (the upper limit of the interval in which the 90-percentile falls), giving us 53.7 as the value of the 90th percentile.

The other percentiles are found in the same manner.

The percentiles (deciles) just calculated are to be interpreted or read as follows:

10% of pupils equal or exceed, and 90% of pupils fall short of, 53.7
(90-percentile)

20% of pupils equal or exceed, and 80% of pupils fall short of, 47.5
(80-percentile)

30% of pupils equal or exceed, and 70% of pupils fall short of, 43.4
(70-percentile)

90% of pupils equal or exceed, and 10% of pupils fall short of, 22.9
(10-percentile)

It should be pointed out that the *median* is the same as the *50th percentile* for all distributions, the median being a special case of a percentile.

The main advantages of percentile norms are:

- (1) They are simple in their interpretation;
- (2) They are applicable in high school subjects where age and grade norms have little meaning;
- (3) Certain percentiles like the deciles divide the pupils into successive levels of accomplishment which are *tenths* (or other simple fractions) of a distribution of abilities of unselected pupils. (Quartiles — i.e., the 75-percentile and the 25-percentile — and the median are often employed, thus breaking the total distribution into quarters.)

The main limitations of percentile norms are:

- (1) Their calculation requires some statistical knowledge, and their meaning is not so simple as age and grade norms;
- (2) They do not permit direct comparisons with mental test results;
- (3) They are somewhat less reliable than averages(means) and *T*-scores or other measures based upon standard deviations.

4. *T-scores.* The use of *T*-scores by this name was introduced by McCall, although the underlying principle

has been in common use for many years. The principle of the *T*-score was implicit in Galton's original approach to the theory of correlation, and has frequently been called to the attention of statisticians by Woodworth, Thorndike, Kelley, and more recently, with some modifications, by McCall. The same principle has also been employed under such varying names as "standard measures," "sigma values," "kentals," and "centigrades."

The *T*-score is based upon the standard deviation as a unit of variability, and hence it has a somewhat smaller probable error than percentiles. It may be defined by the following formula :

$$T = 50 + \frac{10(X - M)}{\sigma}, \text{ where}$$

- T* is the *T*-score
- X* is any raw point-score on a test
- M* is the mean (average) of the distribution of the scores for the group on which the *T*-scores are computed
- σ is the standard deviation (sigma) of the distribution of the scores for the group on which the *T*-scores are computed
- 10 is introduced as an arbitrary constant to eliminate decimals
- 50 is introduced so that the average *T*-score will be 50, and the others will extend above and below this point

The meaning of the *T*-score will be made clearer by an actual calculation of such scores for an obtained distribution. It will first be necessary to calculate *M* and σ . Table 74 shows the preliminary computations, and Table 75 shows the derivation of *T*-scores proper.

The *T*-score has one interesting property; viz., for normal distributions the standard deviation of the distribution of *T*-scores is 10.0. For such distributions, approximately 34 per cent of the *T*-scores fall between the mean (50) and 60. Likewise, about 34 per cent of the *T*-scores fall between 50 and 40. The range of *T*-scores from 40 to 60 would therefore

TABLE 74

SHOWING THE CALCULATION OF THE MEAN (M) AND THE STANDARD DEVIATION (σ) FOR 1036 SCORES ON THE RUCH-POOPENOE GENERAL SCIENCE TEST (DATA FROM TABLE 73)

CLASS INTERVALS	MID-POINTS OR CLASS MARKS	f FREQUENCY	d'	$f d'$	$f d'^2$
77.5-82.5	80	1	8	8	64
72.5-77.5	75	3	7	21	147
67.5-72.5	70	11	6	66	396
62.5-67.5	65	24	5	120	600
57.5-62.5	60	25	4	100	400
52.5-57.5	55	52	3	156	468
47.5-52.5	50	90	2	180	360
42.5-47.5	45	127	1	127	127
37.5-42.5	40	129	0	+ 778	
32.5-37.5	35	153	- 1	- 153	153
27.5-32.5	30	177	- 2	- 354	708
22.5-27.5	25	151	- 3	- 453	1359
17.5-22.5	20	66	- 4	- 264	1056
12.5-17.5	15	25	- 5	- 125	625
7.5-12.5	10	2	- 6	- 12	72
		1036		- 1361 - 1361 + 778 - 583	6535

$$M = 40.0 - \frac{5(1361-778)}{1036} = 37.19$$

$$\sigma = 5 \sqrt{\frac{6535}{1036} - \frac{(583)^2}{(1036)^2}} = 12.24$$

TABLE 75

SHOWING THE CALCULATION OF *T*-SCORES FOR THE SAME DATA
GIVEN IN TABLE 74

X (Raw Scores)	$X - M$ (Deviations of Scores from Mean)	$\frac{10(X - M)}{\sigma}$ (Integral Values)	$50 + \frac{10(X - M)}{\sigma}$ (<i>T</i> -Scores)
80	42.81	35	85
75	37.81	31	81
70	32.81	27	77
65	27.81	23	73
60	22.81	19	69
55	17.81	15	65
50	12.81	10	60
45	7.81	6	56
40	2.81	2	52
37	- 0.19	0	50
35	- 2.19	- 2	48
30	- 7.19	- 6	44
25	- 12.19	- 10	40
20	- 17.19	- 14	36
15	- 22.19	- 18	32
10	- 27.19	- 22	28
5	- 32.19	- 26	24
0	- 37.19	- 30	20

Mean = 37.19

Standard Deviation = 12.24

include roughly two thirds of all the scores. Approximately one sixth of the scores would be above the *T*-score 60, and about the same number would fall below 40.

T-scores can be calculated for any distribution of scores, but it is desirable to have some standard group as a basis. In the elementary school subjects the most often used "standard" group is a distribution of unselected 12-year-old pupils, since that age group is usually the largest numerically and hence is probably the most unselected.

The major advantages of *T*-scores are :

- (1) They are a highly reliable set of measures for norms ;
- (2) Within limits, they can be used with other age, grade, or subject groups than the one on which they were derived ;
- (3) They have wide utility in high school subjects ;
- (4) If mental test scores are first converted to *T*-scores, rather direct comparisons of educational-mental standings can be made.

The main limitations of *T*-scores are :

- (1) Difficulty of calculation and understanding ;
- (2) They are less direct in their comparisons with mental test results expressed in terms of age scales.

CHAPTER TWENTY

THE CONSTRUCTION OF EDUCATIONAL AND MENTAL TESTS (*Continued*)

IV. DETERMINATION OF RELIABILITY OF THE TEST

1. *The Coefficient of Reliability*

Methods of ascertaining reliability. The final step in the experimental work of test construction is the determination of the reliability of the test. This is usually done by correlation methods, and the resulting correlation coefficients are called *reliability coefficients*. A large number of such reliability coefficients, together with certain derived measures, was reported in the chapters of Part II. It is the main purpose of the present chapter to explain the significance and the computation of such measures.

Reliability has previously been defined as the *degree to which a test measures whatever it does measure, regardless of what it may be claimed to measure*. It is an aspect of validity; in fact, it is that aspect of validity which deals with the accuracy of the test as a measuring instrument, and that alone. No matter how carefully the test items are selected, unless the numbers of items be very large indeed, the test may yet prove to be relatively unreliable. A good test is one which entitles the user of the test to place confidence in the scores of the pupils as representative of rather exact quantitative measures of achievement. There is no exact way of ascertaining whether the rank and file of present-day tests sin more against validity or against reliability, but the greater ease of revealing weaknesses in reliability sometimes gives the impression that unreliability is the greater weakness.

There are at least three ways of determining coefficients of reliability; viz.,

1. By repetition of the same test after an interval supposedly great enough to eliminate most of the "memory effect" and yet not long enough for much true growth in ability to take place. This method is the least satisfactory of the three mentioned, and its use should be confined to tests existing in but a single form.
2. By breaking the test into chance halves (usually the odd- and even-numbered items), correlating the half scores, and then "stepping up" the r obtained by the Spearman-Brown formula so as to approximate the r which would be obtained by Method 3 (the correlation of equivalent or similar forms). This method should also be confined to tests existing in but one form, although it may be used with scores from a single form of a test with duplicate forms.

3. By correlations of sets of scores from two different forms of the same test applied to the same group of pupils. This is by far the most trustworthy method.

These three methods will be discussed in some detail.

Method 1. The chief use of this method has been found in certain tests like the Binet-Simon Scale, where Method 3 is impossible since there is but one form of the test. The actual calculations are identical with those of Method 3 and hence need no further discussion here.

Method 2. A large number of elementary and high school tests provide but one form. The general disadvantages of this situation have been pointed out, the calculation of reliabilities being an additional shortcoming of such tests in the light of present purposes.

The detailed steps in determining the reliability of a test provided with but one form will be carried through, using data from a small class of 16 pupils obtained by the use of the Rugg-Clark Standardized Tests in First-Year Algebra. This test consists of two booklets. Booklet 1 contains Tests 1 to 9 and covers the following topics: (1) Collecting Terms,

24 examples; (2) Substitution, 20 examples; (3) Subtraction, 21 examples; (4) Simple Equations, 25 examples; (5) Parentheses, 42 examples; (6) Special Products, 24 examples; (7) Exponents, 36 examples; (8) Factoring, 25 examples; and (9) Clearing of Fractions, 16 examples. There is a total of 233 possible points for Booklet 1, and the working time allowed is 27 minutes. In this particular class, Test 9, Clearing of Fractions, had to be omitted because the class had not yet reached that topic. The test was given about two weeks before the close of the first semester by an experienced test examiner. The scoring was done by the same person.

It should be noted that 16 cases are far too few for the determination of reliability, such a small class being selected in order to simplify the illustration. In estimating the reliability of a test, several hundred pupils should be used for each calculation. If reliability coefficients are calculated grade by grade, or age by age, there should be several hundred pupils in each age or grade, the cases being selected as typical of normal teaching conditions.

The detailed steps are as follows:

- (1) Give the test to a typical group of pupils; one hundred to two hundred are perhaps the minimum numbers needed;
- (2) Score the tests as directed;
- (3) Add separately all scores earned on even-numbered items and on odd-numbered items. The sum of these two half-scores will equal the total score. If there are many separate parts to the test (as is the case with the Rugg-Clark tests), this halving process should be done *within* all parts.
- (4) Correlate the odds and evens by the Pearson product-moment formula,

$$r = \frac{\sum xy}{N\sigma_1\sigma_2}. \text{ Call this } r \text{ the } r_{\frac{1}{2} \frac{1}{2}}$$

(5) Substitute the r found by the odd-even method in the Spearman-Brown formula,

$$r_{nn} = \frac{nr}{1 + (n - 1)r} \quad \text{The resulting}$$

value for r_{nn} will be the desired reliability coefficient.

Table 77 also gives the values for r_{nn} for many values of $\frac{r_1}{z} \frac{r_2}{z}$.

Table 76 shows the calculation of $\frac{r_1}{z} \frac{r_2}{z}$ (odds vs. evens).

The actual formula for r is a variate of the one given above. It is used as a labor-saving device, since it allows the computation of r from an arbitrary mean ("guessed" or assumed mean), with resulting saving of computation.

TABLE 76

CALCULATION OF THE COEFFICIENT OF CORRELATION (RELIABILITY COEFFICIENT) FOR CHANCE HALVES OF THE RUGG-CLARK ALGEBRA TESTS

X (Odds)	Y (Evens)	x'	y'	x'^2	y'^2	$x'y'$
55	56	25	26	625	676	650
19	16	-11	-14	121	196	154
43	42	13	12	169	144	156
13	13	-17	-17	289	289	289
40	34	10	4	100	16	40
43	34	13	4	169	16	52
22	23	-8	-7	64	49	56
54	56	24	26	576	676	624
37	36	7	6	49	36	42
15	20	-15	-10	225	100	150
30	30	0	0	0	0	0
18	14	-12	-16	144	256	192
54	48	24	18	576	324	432
16	20	-14	-10	196	100	140
29	28	-1	-2	1	4	2
7	8	-23	-22	529	484	506
Σ (sums)		15	-2	3883	3366	3485

Assumed means: $M_x = 30$
 $M_y = 30$

With values given in the row of sums at the bottom of Table 76, it is a simple matter to figure the r by the use of the formula :

$$r = \frac{\frac{\Sigma x'y'}{N} - \left(\frac{\Sigma x'}{N} \cdot \frac{\Sigma y'}{N} \right)}{\sqrt{\frac{\Sigma x'^2}{N} - \left(\frac{\Sigma x'}{N} \right)^2} \sqrt{\frac{\Sigma y'^2}{N} - \left(\frac{\Sigma y'}{N} \right)^2}}$$

The actual solution follows :

$$r = \frac{\frac{3485}{16} - \left(\frac{15}{16} \cdot \frac{-2}{16} \right)}{\sqrt{\frac{3833}{16} - \left(\frac{15}{16} \right)^2} \sqrt{\frac{3366}{16} - \left(\frac{-2}{16} \right)^2}} = .97$$

The two radical expressions forming the denominator of the above fraction give the standard deviations of the odds and evens, or :

$$\sigma_x = \sigma_{\text{odds}} = 15.45$$

$$\sigma_y = \sigma_{\text{evens}} = 14.50$$

The true means (in contrast with the arbitrary or assumed means actually used) are found :

$$M_x = 30 + \frac{15}{16} = 30.94$$

$$M_y = 30 + \frac{-2}{16} = 29.87$$

The next step is the estimation of the reliability of the whole test; i.e., the correlation to be expected if the *total* (not the half, or odd and even, scores) were correlated with a *second form* of the Rugg-Clark tests. (There is no second form, in fact.) The expected correlation is obtained by substituting the obtained r (.97, or what has previously been called $r_{\frac{1}{2} \frac{1}{2}}$) in the Spearman-Brown formula, setting N equal to 2 because the length is being *doubled*. Thus :

$$r_{nn} = \frac{Nr}{1 + (N - 1)r} = \frac{(2)(.97)}{1 + (2 - 1)(.97)} = .985.$$

TABLE 77

TABLE FOR OBTAINING DIRECTLY VALUES OF r_{nn} FOR THE SPEARMAN-BROWN PROPHECY FORMULA FOR VARIOUS VALUES OF r AND n

r	n									
	2	3	4	5	6	7	8	9	10	
.10	.18	.25	.31	.36	.40	.44	.47	.50	.53	
.20	.33	.43	.50	.56	.60	.64	.67	.69	.71	
.30	.46	.56	.63	.68	.72	.75	.77	.79	.81	
.40	.57	.67	.73	.77	.80	.82	.84	.86	.87	
.50	.67	.75	.80	.83	.86	.88	.89	.90	.91	
.60	.75	.82	.86	.88	.90	.91	.92	.93	.94	
.70	.82	.87	.90	.92	.93	.94	.95	.95	.96	
.80	.89	.92	.94	.95	.96	.96	.97	.97	.98	
.90	.947	.964	.973	.978	.981	.984	.986	.988	.989	
.91	.953	.968	.976	.981	.984	.986	.988	.989	.990	
.92	.958	.972	.979	.983	.986	.988	.989	.990	.991	
.93	.964	.976	.982	.985	.988	.989	.991	.992	.993	
.94	.969	.979	.984	.987	.989	.991	.992	.993	.994	
.95	.974	.983	.987	.990	.991	.993	.993	.994	.995	
.96	.980	.986	.990	.992	.993	.994	.995	.995	.996	
.97	.985	.990	.992	.994	.995	.996	.996	.997	.997	
.98	.990	.993	.995	.996	.997	.997	.997	.998	.998	
.99	.995	.997	.997	.998	.998	.999	.999	.999	.999	

The value, .985, then, is the best estimate which we can obtain for the reliability coefficient of the *entire* Rugg-Clark Algebra Tests.

The reader should note that, even with as large an r as .97, there are some striking differences in the scores earned on the two halves of the test. Reference to Table 76 shows that the divergences range from 0 to 9 (the sixth pair of scores from the top of the table), a disagreement of about $\frac{9}{15}$ or $\frac{3}{5}$ of a standard deviation.

The obtained reliability coefficient (.985) is very high for such a small group of pupils and a single class. Very few educational tests will prove this reliable, as Part II of this volume has shown.

Table 77 above gives a short list of values for the Spear-

man-Brown formula. For the values of r and n shown, estimated coefficients may be read directly from this table.

The use of the entries for n is identical with the example just given; viz., if the test is doubled (as in "stepping up" half-forms to whole forms), n is taken as 2.

Method 3. This is by far the simplest procedure when two or more forms have been given to the same pupils.

A new set of data will be used here in presenting the third method in order that a more convenient set of values for use in a later section can be secured. Certain data from Chapter IX of Part II on the Barr Diagnostic Tests in American History will serve our purposes as well as any. There are two forms of the Barr tests, designated as Series 2 A and Series 2 B. These were given to 279 twelfth-grade pupils at intervals of two days between forms. Since the treatment of the entire 279 pairs of scores would require more space than could be justified, a sampling of 25 pairs of scores was drawn from the complete tabulation by taking every eleventh pair in the list until 25 pairs had been drawn.

This raises a question of sampling which may be worth a few comments in passing. First, the original lot and the sampling may be compared as a pure matter of sampling:

	MEAN SCORES		STANDARD DEVIATIONS		r
	2 A	2 B	2 A	2 B	
Original lot (279)	47.5	48.5	12	12	.71 \pm .02
Sampling (25)	44.4	44.7	14.3	12.8	.84 \pm .04

The r obtained by the sampling is more than three times its probable error larger than the r on the entire lot. The only significant comment on this difference is that it shows the inadequacy of small samplings. Fortunately, no injustice can be done to the Barr tests by means of our illustration, since the sample chosen is probably more reliable

than can be expected with repeated samplings of the size of typical classes. Table 78 shows the calculation of the reliability coefficient based on the sample.

TABLE 78

THE CALCULATION OF THE RELIABILITY COEFFICIENT BY CORRELATION OF TWO EQUIVALENT FORMS OF A TEST. THE VARIABLES ARE SCORES ON THE BARR DIAGNOSTIC TESTS IN AMERICAN HISTORY.

X (Series 2A)	Y (Series 2B)	x'	y'	x'^2	y'^2	$x'y'$
45	37	0	- 8	0	64	0
26	23	- 19	- 22	361	484	418
39	52	- 6	7	36	49	- 42
45	39	0	- 6	0	36	0
29	41	- 16	- 4	256	16	64
38	26	- 7	- 19	49	361	133
38	42	- 7	- 3	49	9	21
29	32	- 16	- 13	256	169	208
50	50	5	5	25	25	25
70	53	25	8	625	64	200
30	33	- 15	- 12	225	144	180
59	57	14	12	196	144	168
65	75	20	30	400	900	600
55	53	10	8	100	64	80
41	48	- 4	3	16	9	- 12
73	69	28	24	784	576	672
35	34	- 10	- 11	100	121	110
24	32	- 21	- 13	441	169	273
66	60	21	15	441	225	315
46	56	1	11	1	121	11
33	31	- 12	- 14	144	196	168
37	37	- 8	- 8	64	64	64
56	50	11	5	121	25	55
27	40	- 18	- 5	324	25	90
55	48	10	3	100	9	30
Σ (sums)		- 14	- 7	5114	4069	3885
						- 54
						3831

Assumed means:

True means:

Sigmas:

$M_x = 45$

$M_x = 44.4$

$\sigma_x = 14.3$

$M_y = 45$

$M_y = 44.7$

$\sigma_y = 12.8$

$r = .84 \pm .04$

The true means, standard deviations, and the coefficient of correlation were computed by the same methods as were shown for the preceding example (Rugg-Clark Algebra Tests).

The r obtained (.84) gives the reliability of one (either) form directly.

In the next section the question of the value of the reliability coefficient as a measure of reliability will be discussed, together with certain derived measures which have special advantages over the reliability coefficient.

2. Measures of Errors in Individual Scores

The reliability coefficient an unanalyzed measure. The reliability coefficient, *per se*, is not a very satisfactory measure of the accuracy of a test, for several reasons:

- (1) It is a generalized and unanalyzed measure of the trends of behavior in large numbers of scores;
- (2) It tells little or nothing about the margin of error in the score of an individual pupil (which is the important thing);
- (3) By itself — i.e., unsupported by other facts — it has little meaning because reliability r 's fluctuate with changes in range of talent (the extent of individual differences among the pupils on whom the r is computed).

The question of *range of talent* in its effect upon the size of the reliability coefficient may be discussed first. Kelley¹ has expressed the relation between the magnitude of the obtained reliability coefficient and the magnitude of the standard deviation (which is the best measure of range of talent, generally) by the following formula:

$$\frac{\sigma_1}{\Sigma_1} = \frac{\sqrt{1 - R}}{\sqrt{1 - r}}$$

¹ Kelley, T. L., "The Reliability of Test Scores." *Journal of Educational Research* (May, 1921).

Where

- σ_1 is the smaller range of talent (standard deviation)
- Σ_1 is the larger range of talent (standard deviation)
- r is the reliability coefficient figured for the smaller range of talent
- R is the *expected* reliability coefficient figured for the larger range of talent (upon the assumption that the test is equally effective — i.e., reliable — in both the wide and the narrow range of talent)

The usefulness of this formula can be made clear by an example.¹

Examiner A reports a reliability of 0.95 for Test X. Examiner B reports a reliability of 0.64 for the same test. The complete data from the two investigations are as follows :

EXAMINER	r	N	σ	GROUP TESTED
A	0.95	500	30.3	All grades; IV to XII
B	0.64	500	10.1	Grade VI only

Substituting these data in the formula just given, and solving for R , we get

$$\frac{\sigma_1}{\Sigma_1} = \frac{\sqrt{1 - R}}{\sqrt{1 - r}} ; \quad \frac{10.1}{30.3} = \frac{\sqrt{1 - R}}{\sqrt{1 - .64}} ; \text{ and } R = .96.$$

If Examiner A had drawn the conclusion (as test workers often do) that Examiner B's results were in error (too low), or vice versa, this conclusion would have been misleading. As a matter of fact, if we make due allowance for the differences in the ranges of talent by means of Kelley's formula,

¹ See Ruch, G. M., "Minimum Essentials in Reporting Data on Standard Tests," *Journal of Educational Research*, Vol. XII, No. 5 (1925), pages 349-358, for a general discussion of the topics covered in this chapter.

the two findings are in substantial agreement. Examiner B found an r of .64 for a small range of talent ($\sigma = 10.1$). Had he used a range three times as large (as did Examiner A), it has been estimated that he would have obtained an r in the neighborhood of .96 (almost exactly the value that A reported).

This illustration shows the inadequacy of the mere statement of reliability coefficients *alone* as evidence of the reliability of a test. A number of other facts are essential or highly desirable, as follows:

- (1) The reliability coefficient (together with a designation of the method employed; e.g., repetition of the same form, correlation of chance halves, or correlation of similar forms);
- (2) The standard deviations of the test form or forms (as a measure of range of talent);
- (3) The mean (average) scores on all forms of the test used (These are not useful directly for purposes of proving reliability, but are valuable as a check on equivalence of forms and for certain other calculations);
- (4) The numbers of cases used;
- (5) A description of the talent used (range of ages or grades, or other data bearing on the representativeness of the sampling).

These data will permit the calculation of almost every statistical measure needed in critical test evaluation; e.g., *true scores*, *probable errors of test scores*, *T-scores*, etc.

In elementary school tests, reliability coefficients are often reported for *unselected age groups*. In such cases it is not so necessary, although it is still desirable, to publish the standard deviations, since such age groups have by general agreement come to be regarded as "standard groups" for thinking about reliability. It is fairly easy to duplicate age groups

in such a way as to guarantee practical equality of ranges of talent. The unselected grade group is a somewhat less stable standard for comparisons.

The probable error of a test score. There are a number of methods of expressing the amount of error present in an individual score, two of which will be discussed here; viz., (a) *the probable error of a raw score* and (b) *the probable error of an estimated true score*. Before entering upon a discussion of these two methods, it will be well to define the terms "raw score" and "true score."

1. The raw score on a test is the obtained point-score; i.e., the score given the pupil after the correction of the test. Raw scores are what are termed merely "scores" on a test. The scores treated in the two tables showing correlation in this chapter are raw scores.

2. The true score is a hypothetical concept. We have seen from Tables 76 and 78 that the scores of individual pupils fluctuate from form to form of a test. This is due to the fact that each form of the test is merely a sample of the pupil's ability. If a single form of a test were made, successively, two, three, four, five, etc., times as long by the addition of similar items, the error due to limited sampling would grow relatively smaller and smaller. If the test were made infinitely long, the error in each individual score would be zero, because the measurement would be complete and no sampling errors would remain. The true score can therefore be defined as follows:

$$X_{\infty} = \frac{X_1 + X_2 + X_3 + X_4 + \dots + X_n}{n}; \text{ where}$$

X_{∞} is the symbol adopted for a true score

X_1 is the score on the first form of the test

X_2 is the score on the second form of the test, etc.

The true score thus defined becomes the average score on an infinite number of equivalent or "similar" tests (equivalent or similar implying (a) homogeneous samplings, (b) equal means, and (c) equal standard deviations).

The formula for the probable error of a *raw score* may be stated:

$$\text{P.E.}_{(\text{raw score})} = .6745 \sigma_1 \sqrt{1 - r_{12}}$$

The formula for the probable error of an *estimated true score* may be stated:

$$\text{P.E.}_{(\text{estimated true score})} = .6745 \sigma_1 \sqrt{r_{12} - r_{12}^2}$$

In these formulas, r_{12} is the reliability coefficient of the test as found by correlating one form against a second form.

In actual practice σ_1 is taken as $\frac{\sigma_1 + \sigma_2}{2}$; i.e., the average of the two standard deviations.

Returning to the data of Table 78 (the Barr history tests), we find the following computations needed in computing the probable error of a raw score:

r_{12} (the correlation of Series 2 A and 2 B)	0.84
σ_1 (the standard deviation of Series 2 A)	14.3
σ_2 (the standard deviation of Series 2 B)	12.8

Substituting in the formula,

$$\begin{aligned} \text{P.E.}_{(\text{raw score})} &= .6745 \sigma_1 \sqrt{1 - r_{12}}, \text{ we get} \\ &= .6745 \frac{14.3 + 12.8}{2} \sqrt{1 - .84} = 3.7 \end{aligned}$$

The meaning of the 3.7 thus obtained is very simple. If we take a given score, — e.g., 60, — we can say:¹

¹ These statements are not strictly true, because the margin of error is different with raw scores at different parts of the range. Very high scores and very low scores are more subject to errors, due to what are termed "regression effects." The significance of this footnote will be made clearer by the later discussion of the probable error of an estimated true score.

The chances are $1:1$ that the true score of this pupil lies between $60 - 3.7$ and $60 + 3.7$; i.e., between 56.3 and 63.7;

The chances are about $4:1$ that the true score lies between $60 - 2(3.7)$ and $60 + 2(3.7)$; i.e., between 52.6 and 67.4;

The chances are about $20:1$ that the true score lies between $60 - 3(3.7)$ and $60 + 3(3.7)$; i.e., between 48.9 and 71.1.

Even now, with the probable error of a raw score estimated at 3.7, we cannot visualize exactly the significance of the finding. Barr gives as temporary standards the following:

TEST	I	II	III	IV	V	TOTAL
Series 2 B (Grade VIII)	7.2	7.1	5.7	9.6	5.4	35.0
Series 2 B (Grade XII)	10.0	8.0	9.0	12.3	7.5	46.8

The difference in the total scores between Grade VIII and high school is $46.8 - 35.0$, or 11.8 points. The probable error of the raw score figured in ratio to the differences between Grade VIII and Grade XII is $3.7/11.8$, or slightly more than a third. We can now restate the probabilities given before, as follows:

The chances are about $1:1$ that the obtained score differs from the true score by an amount not greater than roughly $\frac{1}{3}$ the difference between eighth-grade and twelfth-grade accomplishment;

The chances are about $4:1$ that the obtained score differs from the true score by an amount not greater than roughly $\frac{2}{3}$ the difference between eighth-grade and twelfth-grade accomplishment; etc.

There is yet another way of looking at the P.E. of the raw score; viz., in relation to the size of the standard deviation. The average of the standard deviations of the two forms was found to be about 13.5. The ratio needed is $\frac{3.7}{13.5}$, or .27.

A series of statements about the probabilities of errors expressed as fractions of the standard deviation could be drawn up by analogy to those before, the first one being:

The chances are about 1:1 that the obtained score differs from the true score by an amount not greater than roughly .25 of the standard deviation of the entire distribution; etc.

The probable error of an estimated true score will next be considered.

It has already been pointed out that the true score is a hypothetical concept. We can, however, obtain an *estimated* true score for each pupil if we know the following facts:

- (1) The reliability coefficient of the test;
- (2) The average score of a group of pupils on the test (the same group on which r was figured).

We have these data for the Barr tests (Table 78). The formula needed for estimating true scores is given by Kelley as follows:¹

$\bar{X}_\infty = r_{12} X_1 + (1 - r_{12}) M_1$. The notation of Kelley has been changed to fit our previous use of symbols, as follows:

\bar{X}_∞ is the estimated true score to be found

r_{12} is the reliability coefficient (.84 in our problem)

X_1 is each raw score, taken in turn

M_1 is the mean of the distribution of raw scores (44.4)

¹ Kelley, T. L., *Statistical Method* (The Macmillan Company, 1923), pages 214-216.

In Table 78 the scores for Series 2A and 2B of the Barr tests were given. An estimated true score could be found for each raw score on each form. We shall confine the present discussion to the raw scores on Series 2A (the values headed X in Table 78).

TABLE 79

SHOWING THE CALCULATION OF ESTIMATED TRUE SCORES FOR
SERIES 2A OF THE BARR DATA FROM TABLE 78

X_1	$.84 X_1$	$8.4 X_1 + 7.1$ or $(\bar{X})_{\infty}$
45	37.8	44.9
26	21.8	28.9
39	32.8	39.9
45	37.8	44.9
29	24.4	31.5
38	31.9	39.0
38	31.9	39.0
29	24.4	31.5
50	42.0	49.1
70	58.8	65.9
30	25.2	32.3
59	49.6	56.7
65	54.6	61.7
55	46.2	53.3
41	34.4	41.5
73	61.3	68.4
35	29.4	36.5
24	20.2	27.3
66	55.4	62.5
46	38.6	45.7
33	27.7	34.8
37	31.1	38.2
56	47.0	54.1
27	22.7	29.8
55	46.2	53.3
(M) 44.4		44.4

Substituting in the formula, we have

$$\begin{aligned}\bar{X}_{\infty} &= .84 X_1 + (1 - .84)44.4 \\ &= .84 X_1 + 7.1\end{aligned}$$

Table 79 carries the scores of Form 2 A of the Barr tests through the steps leading to the estimated true scores by the use of the formula just given.

Attention must be called to one interesting property of these estimated true scores; viz., the estimated true scores differ most from the raw scores for large and small values (large and small distances from the average). Near the mean the scores change very little, and at the mean they do not change at all.

The following three scores have been selected as the highest, the nearest the mean, and the lowest raw scores, respectively. Their estimated true scores are shown, together with the differences:

RAW SCORE (X_1)	EST. TRUE SCORE (X_{∞})	DIFFERENCE
73	68.4	4.6
45	44.9	0.1
24	27.3	3.3
Maximum Score 100	91.1	8.9
Minimum Score 0	7.1	7.1

The principle to be drawn from the behavior of the raw scores in comparison with the best estimate of the true scores may be stated as follows: *Very high obtained raw scores as a matter of unreliability are more likely to be in error upward; i.e., to be too high. Very low obtained raw scores are more likely to be in error downward; i.e., to be too low.* An even more compact statement is that: *Unreliable test scores regress toward the mean from both directions as the scores*

approach the true scores. This is often called the principle of *regression*, and was discovered empirically a long time ago by Sir Francis Galton while studying the relation between the heights of parents and offspring.

The practical lesson from the foregoing account is that we cannot place the same confidence in extremely high scores that we can in scores nearer the average. The same is true for very low scores. It will also be recalled that in a footnote on page 367 the statement was made that the probable error of a raw score cannot be stated as exactly as the probable error of an estimated true score. The reason for this statement lies in the inequality of error, both in amount and in direction, in different parts of the scoring scale. This fact has been shown for the Barr tests.

The illustration just chosen shows the behavior of a relatively reliable test. Had another test, say with a reliability of .50 (and there are many such) and a mean of 44.4, been selected, the estimated true scores corresponding to 73, 45, and 24, respectively, would prove to be 58.7, 44.7, and 34.2, respectively, and the changes would be 14.3, 0.3, and 10.2 score points, in turn.

With tests having reliabilities less than .90, significant changes in the raw scores in comparison with estimated true scores will be found for many pupils. For tests of very low reliability (e.g., below .75), it will be wise to compute the estimated true scores and use these instead of raw scores for all important purposes, such as grade classification or experimental investigations.

With the meaning of estimated true scores in mind, we can turn to the question of the probable errors of such scores.

The formula of the probable error of an estimated true score has already been given. Substituting the data for the Barr tests, we have

$$\begin{aligned}\text{P.E.}(\text{estimated true score}) &= .6745 \frac{\sigma_1 + \sigma_2}{2} \sqrt{r_{12} - r_{12}^2} \\ &= .6745 \frac{14.3 + 12.8}{2} \sqrt{.84 - (.84)^2} \\ &= 3.4\end{aligned}$$

The interpretation of this probable error is the same as that previously given for the probable error of a raw score, except, of course, that it must be used in connection with \bar{X}_∞ values (estimated true scores). It is to be preferred to the probable error of the raw score, since it holds equally well for any score, high or low.

There is one remaining issue of considerable moment in arriving at a valid expression of reliability; viz., the comparison of probable errors of scores on different tests *where the test units are not the same*. A score of 50 on the Otis General Intelligence Examination has a very different meaning from the score of 50 on the Barr Diagnostic Tests in American History, for many reasons:

- (1) They are not scaled from the same zero-point;
- (2) They do not have the same mean scores for a given group;
- (3) Their standard deviations are different.

For practical purposes the mean and the standard deviation of a test fix the meaning of the score units. We can equate the scores on two different tests if both have been given to the same pupils and the means and standard deviations determined. Kelley¹ has suggested that in comparing the probable errors of tests scaled to different units the following ratio should be used:

$$\frac{\text{P.E.}(\text{estimated true score})}{\sigma_1}, \text{ or, more simply, } \frac{\text{P.E.}_\infty}{\sigma_1}.$$

¹ *Statistical Method* (The Macmillan Company, 1923), pages 214-216.

For the scores on the Barr 2A test this ratio equals $\frac{3.4}{14.3}$, or roughly 24 per cent. The meaning of this ratio (.24) is the same as for the probable error of any measure; i.e., the chances are 1:1 that the true score does not differ from the estimated true score more than $.24\sigma$. The chances are about 4:1 that the true score does not differ from the estimated true score more than $2(.24\sigma)$; i.e., $.48\sigma$. The chances against the error being as great as $.72\sigma$ are about 20:1. The chances against the error being as great as 1σ are at least 150:1.¹

Following the same reasoning as in the case of the probable error of a raw score, the ratio of the $P.E_{\infty,1}$ to the difference between the eighth-grade and twelfth-grade norms (11.8 points) is roughly 0.3. Approximately half of the pupils would be located with an error no larger than three tenths of the difference between eighth-grade and twelfth-grade accomplishment.

The foregoing treatment of the reliability of the test completes the actual experimental work on the test. The remaining work to be done falls under the caption of "Perfecting the Administration of the Test."

V. PERFECTING THE ADMINISTRATION OF THE TEST

The Manual of Directions. It will be unnecessary to repeat the detailed content of the Manual of Directions or Examiner's Guide, as it is variously called, after the thorough

¹ This reasoning assumes that a sufficiently large sampling has been used to establish the constants used in arriving at the estimated true scores and their errors. The twenty-five sets of scores are far too small a number to justify any confidence in the values actually quoted. These are illustrative merely. Indeed, the larger population of 279 pupils from which our sample was drawn yielded a considerably lower r ; viz., .71. Using the value 12, the standard deviation of the population of 279, and r as .71, the probable error of an estimated true score is 3.67, or about .33 of a standard deviation.

discussion given in this and the three preceding chapters. It will be sufficient to list the topics that a well-worked-out Manual should contain:

- (1) An account of the purpose of the test;
- (2) A treatment of the validation of the test;
- (3) Directions for administering the test;
- (4) Directions for scoring;
- (5) The tables of norms, and directions for their use in interpreting the results of the test;
- (6) The facts about the reliability of the test;
- (7) Suggestions for the uses of the results, the remedial program, diagnosis, etc.

The scoring keys. The correct answers should be provided with the test. These are most conveniently used if placed on cardboard strips in such a way that they can be cut apart and superimposed directly on the pupils' test blanks. It is usually unwise to depend entirely upon answer lists in the body of the Manual, unless the test is a very short one. Directions for scoring doubtful responses must be given. If corrections for chance are employed, these should be explained fully.

Selected References for Part IV

The following books treat the topic of test construction somewhat more at length than do Chapters XVII to XX:

MC CALL, W. A. *How to Measure in Education*. The Macmillan Company, New York; 1922.

MONROE, W. S. *Introduction to the Theory of Educational Measurements*. Houghton Mifflin Company, Boston; 1923.

OTIS, ARTHUR S. *Statistical Method in Educational Measurement*. World Book Company, Yonkers-on-Hudson, New York; 1925.

TERMAN, L. M., et al. *The Stanford Revision and Extension of the Binet-Simon Scale for Measuring Intelligence*. Warwick & York, Inc., Baltimore; 1916.

INDEX

Administration of tests, importance of ease of, 56-58
Age norms, derivation and significance of, 345-347
Algebra tests, descriptions of, 72-84; intercorrelations, 81-82
Aptitude tests, descriptions of, 39-40
Ayres Scale for Measuring Handwriting, 238
Ayres Spelling Scale, 232
Barr Diagnostic Tests in American History, 178-179
Binet-Simon tests, 214-215
Biology tests, descriptions of, 141-145
Blackstone Stenographic Proficiency Tests, 191-192
Botany tests, descriptions of, 141
Briggs English Form Test, 231
Brown-Woody Civics Test, 182
Buckingham Scale for Problems in Arithmetic, Division III, 223-229
Buckingham-Stevenson Place Geography Tests, 235
Chance and guessing in objective tests, 282-294
Chapman-Cook Speed of Reading Test, 232
Charters Diagnostic Language and Grammar Tests, 232
Chemistry tests, descriptions of, 146-152
Classification of pupils, 32-38, 40-42
Coefficient of reliability, 355-363
Columbia Research Bureau Tests, 90, 161-162, 209-210
Commercial tests, descriptions of, 191-192
Compass Diagnostic Tests in Arithmetic, 229-230
Composition scales, descriptions of, 123-128; remedial procedures based upon, 128-130
Construction of educational and mental tests, described in detail, 301-375
Cooprider Information Items in Biology, 144-145
Corrections for chance and guessing, validity of, 282-294
Courtis Standard Practice Tests in Handwriting, 238
Courtis Standard Research Tests, Series B, 230
Courtis Supervisory Geography Test, 235
Criteria for selecting tests, 45-68
Cross English Test, 102-103
Diagnosis of teaching efficiency, 14-16
Diagnostic tests, functions of, 64-65; requirements for, 18-27
Douglass Algebra Tests, 77-80
Duplicate of equivalent forms of tests, need for, 65; equivalence of, 65-66
Dvorak General Science Scales, 140-141
Educational tests, first, 2; early history of, 4; limitations in the past, 5-7
English tests, descriptions of, 97-135; bibliography of, 131-135; in junior high school, 231-235
Equivalence of test forms, how determined, 338-339
Examinations, limitations, 252-265
Experimental try-outs of test items, 329-332
Foreign language tests, descriptions of, 160-176; remedial procedures based upon, 172; bibliography of, 173-176
Franseen Diagnostic Tests in Language, 232

Freeman Chart for Diagnosing Faults in Handwriting, 238

French and Spanish tests, 160-166

Gates-Strang Health Knowledge Test, 238

General science tests, descriptions of, 137-141

Geography tests, descriptions of, 235-237

Geometry tests, descriptions of, 84-91

Glenn-Welton chemistry examinations, 151-152

Godsey Latin Composition Test, 170

Grade norms, 344-345

Gray Standardized Oral Reading Check Tests, 232

Gregory-Spencer Geography Tests, 235

Gregory Tests in American History, Test III, 180-181

Guessing in tests, corrections for, 282-294

Haggerty Intelligence Examination, Delta 2, 216, 218, 219, 221

Haggerty Reading Examination, Sigma 3, 113-115

Hahn-Lackey Geography Scale, 235

Handschin Modern Language Tests, 161

Hawkes-Wood Plane Geometry Examination, 90

Henmon French Tests, 160-161

Henmon Latin Tests, 166-167

Herring Revision of Binet-Simon Tests, 212, 214-215

History tests, descriptions of, 177-184

Holley Sentence Vocabulary Scale, 121

Home economics tests, 238

Hotz Algebra Scales, 72-77

Hudelson English Composition Scales, 123-127

Hughes Physics Scales, 153

Illinois Examination, 239

Illinois General Intelligence Scale, 241

Illinois Standardized Algebra Tests, 80

Informal or objective tests, bibliography of, 295-297; descriptions of, 250-297; limitations and advantages of, 270-274; samples of, 272-281; scoring of, 282-294; types of, 266-270

Intelligence tests, as a basis for classification, 34-35, 40-42; bibliography of, 223-226; comparative data on, 216-217; group tests, 216-226; individual tests, 214-215; in predicting high school marks, 221-223; in junior high schools, 241; limitations of, 223; rise of, 2-3

Iowa Comprehension Tests, 201-202

Iowa Dictation Exercise and Spelling Test, 232

Iowa High School Content Examination, 202-204

Iowa Physics Test, 152-153

Iowa Placement Examinations, in chemistry, 149-151; in English, 106-108; general description of, 2, 3, 209; in French, 162-165; in mathematics, 82-84; in physics, 153-154; in Spanish, 163

Junior high school tests, bibliography of, 242-247; descriptions of, 227-247

Kelley Mathematical Values Test, 92

Kepner Background Test in Social Sciences, 181

King-Clark Foods Test, 239

Kirby Grammar Test, 98-101

Kwalwasser-Ruch Test of Musical Accomplishment, 195-197

Language and grammar tests, descriptions of, 97-108; intercorrelations among, 105-106; remedial procedures based upon, 107-108

Latin tests, 166-172

Lewis Composition Scales, 127

Limited sampling, as a weakness of examinations, 257-265

Lippincott-Chapman Classroom Products Survey Test, 239

Mathematics tests, bibliography of, 93-96; descriptions of, 71-96; in junior high schools, 228-231

Mechanical aptitude tests, 186-188

Mechanical features of tests, importance of, 68

Michigan Botany Test, 141

Miller Mental Ability Test, 216, 218, 219, 221

Minnick Geometry Tests, 84-88

Monroe General Survey Scales in Arithmetic, 230

Monroe Standardized Silent Reading Tests, Test III, 120-121, 233

Monroe Timed Sentence Spelling Tests, 110-111

Motivation of learning, uses of tests in, 43-44

Murdoch Sewing Scale, 239

Music tests, descriptions of, 192-196

National Intelligence Tests, 241

New-type examinations, advantages and limitations of, 270-272; compared with traditional examinations, 251-265; samples of, 272-281; scoring of, 282-294; types of, 266-270

Norms, age, 345-347; grade, 344-345; in the interpretation of test results, 60-62; kinds of, 62-64, 343-354; percentile, 347-350; statistical derivation of, 343-354; *T*-scores, 350-354

Objectification of pupil records, uses of tests for the, 18

Objective tests, advantages and limitations of, 270-272; descriptions of, 251-265; samples of, 272-281; scoring of, 282-294; types of, 266-270

Objectivity, as an essential of a good test, 58-59

Orleans-Solomon Latin Prognosis Test, 171

Otis Arithmetic Reasoning Test, 230

Otis Classification Test, 239

Otis Group Intelligence Scale: Advanced Examination, 216, 218, 219, 220

Otis Scale for Rating Tests, 48, 49

Percentile norms, meaning and calculation of, 347-350

Physics tests, descriptions of, 152-156

Posey-Van Wagenen Geography Scales, 235-236

Powers General Chemistry Test, 148-149

Pressey Diagnostic Tests in English Composition, 103-104

Pressey-Richards American History Test, 180

Pressey Technical Vocabularys of the Public School Subjects, 120-121

Probable error of test scores, 363-374

Prognosis tests, descriptions of, 39-40

Promotions, uses of tests for, 28-30

Pupil progress, measurement of, 10-14

Rating scales, discussion and criticism of, 312-317

Reading tests, descriptions of, 112-123

Reliability, defined and discussed, 51-56; of individual scores, 363-374; of ordinary school examinations, 254-265; statistical determination of, 355-374

Rich Chemistry Tests, 151

Rogers Test of Mathematical Ability, 91-92

Ruch-Cossmann Biology Test, 63, 142-144

Ruch-Popenoe General Science Test, 137-140, 347, 348

Scaling test items, 332-338

Schorling-Sanford Achievement Test in Plane Geometry, 89-90

Science tests, bibliography, 157-159; descriptions of, 136-159; remedial procedures based upon, 145-146

Seashore Measures of Musical Talent, 192-194

Sectioning of classes, 30-34

Seven S Spelling Scales, 109-110

Social studies, tests in, bibliography of, 184-185; descriptions of, 177-184; remedial procedures based upon, 182

Spanish and French tests, 160-166

Spearman-Brown formula, 358-361; table of, 360

Speed *vs.* power in tests, 341-342

Spelling tests, descriptions of, 108-112; remedial procedures based upon, 111-112

Spencer Diagnostic Tests in Arithmetic, 230

Standards of pupil performance, uses of tests for, 16-18

Standardization of tests, described in detail, 301-375

Stanford Achievement Test, 240-241

Stanford Arithmetic Examination, 37, 230-231

Stanford Dictation (Spelling) Test, 232

Stanford Reading Examination, 233

Stanford Revision of Binet-Simon Tests, 212, 214-215

Status of high school measurement, 1-7

Stenquist Mechanical Aptitude Tests, 186-188

Stevenson, Pressey, Tyler, et al., Latin Tests, 170-172

Stone Narrative Reading Tests, 233

Subjectivity in relation to examinations, 254-257

Survey tests, bibliography of, 210-211; descriptions of, 200-211; in the junior high school, 239-241; remedial procedures based upon, 210

Teachers' marks, unreliability of, 254-265

Terman Group Test of Mental Ability, 216, 218, 220, 221

Test construction, detailed methods of, 301-375

Thorndike Handwriting Scale, 238

Thorndike-McCall Reading Scale, 115-118

Thorndike Test of Word Knowledge, 233

Thurstone Vocational Guidance Tests, in algebra, 81; general description of, 188-189; in geometry, 90; in physics, 155

Time limits for tests, how determined, 339-341; resulting effects of, 341-342

Time requirements in testing, 66-68

Trade tests, 189-191

Tressler Minimum Essentials Test (English), 105

True scores, 369-372

T-scores, meaning and calculation of, 350-354

Ullman-Kirby Latin Comprehension Test, 168-169

Unstandardized tests, compared with standard tests, 251-252

Upton-Chassell Citizenship Test, 238

Uses and limitations of tests, in classification of pupils, 36-42; in diagnosis of special difficulties, 18-27; in grading, promotion, and sectioning, 28-34; in school administration and supervision, 8-18; for the motivation of learning, 43-44; for prognosis and aptitudes, 39; for research purposes, 42-43

Validation of educational and mental tests, 301-328

Validity, defined, 48-51, 302-304

Van Wagenen English Composition Scales, 127

Van Wagenen Reading Scales, in American history, 181; in English literature, 118-120; in general science, 141

Vocational subjects, bibliography of tests in, 197-199; descriptions of tests in, 186-199; tests in, for junior high schools, 238-239

Wakefield Diagnostic English Tests, 105

Weighting test items, 332-338

White Latin Test, 169

Wilkins Prognosis Test in Modern Languages, 163-164

Willing Scale for Measuring Written Compositions, 235

Wilson Language Error Test, 101-102

Wisconsin Inventory Test, 231

Woody Arithmetic Scales, 231

Woody-McCall Mixed Fundamentals Test in Arithmetic, 231

*Especially designed for use in high schools
(Also usable as low as Grade 6 and as high as first year in college)*

TERMAN GROUP TEST *of* MENTAL ABILITY

By LEWIS M. TERMAN

Professor of Educational Psychology, Stanford University; joint author of the National Intelligence Tests and of the army mental tests; author of the Stanford Revision of the Binet-Simon Scale, and of a number of books on the measurement of intelligence

This test is unique in many respects. Each of its 886 items was measured against a composite outside criterion. A try-out resulted in a reduction to 370 items, each helping to differentiate bright pupils from dull ones. The items retained are more highly selected than will be found in any other group mental test.

The Terman Test is an eleven-page booklet. The pupil does no writing. The backs of the Scoring Keys contain the scoring rules. Only 30 to 35 minutes will be required to test a group with it. The procedure has been so simplified that it can be mastered by any teacher in a few minutes. The size of the booklets makes their use without desks easy.

Examination: Form A. Price per package of 25 booklets, including Scoring Key and Manual of Directions, \$1.20 net.

Examination: Form B. Price per package of 25 booklets, including Scoring Key and Manual of Directions, \$1.20 net.

Specimen Set. Price 15 cents postpaid.

WORLD BOOK COMPANY

YONKERS-ON-HUDSON, NEW YORK
2126 PRAIRIE AVENUE, CHICAGO

Ruch-Popenoe General Science Test

BY GILES M. RUCH
AND HERBERT F. POPENOE

A RELIABLE and time-saving guide in assigning marks, determining promotions, classifying students according to ability, and advising students in choice of courses. It is designed primarily to measure accomplishment in general science courses of any type in grades 7 to 9, and the purely objective and standardized scores obtained from it furnish valuable and significant data for instructional and administrative purposes.

The test, based upon careful analysis of existing textbooks, samples a wide range of simple knowledge of physics, chemistry, astronomy, physiography, and biological science. It is composed of questions concerning facts, principles, definitions, and applications and there are also diagrams and drawings with exercises.

The test is easy to give and interpret and all necessary instructions are supplied with each package of tests. It can be completed within a 45 minute period and only a short time is required for scoring.

Examination: Form A. Price per package of 25 examination booklets, including Manual of Directions, (Revised), Key, and Class Record, \$1.30 net.

Examination: Form B. Price per package of 25 examination booklets, including Manual of Directions, (Revised), Key, and Class Record, \$1.30 net.

Specimen Set. Price 20 cents postpaid.

WORLD BOOK COMPANY
YONKERS-ON-HUDSON, NEW YORK
2126 PRAIRIE AVENUE, CHICAGO

MILLER MENTAL ABILITY TEST

FOR GRADES 7 TO 12, AND FOR COLLEGE FRESHMEN

By W. S. MILLER

Professor of Educational Psychology, University of Minnesota

THIS test is the result of six years' experimentation in individual and group examination of high school students. Its value as a basis for predicting success in high school work has been demonstrated.

The examination consists of three parts, each containing forty items: A Disarranged Sentences-Directions Test, a Control-Association Test, and an Analogies Test.

The Manual of Directions gives complete instructions for administering and scoring the test and interpreting results. The Key is so arranged as to be laid upon the test paper to aid in rapid scoring. Sheets for making age-grade-score distributions and for plotting percentile graphs are supplied in order to facilitate the interpretation of results.

The test can be given within thirty minutes. The scoring is simple and thoroughly objective. Age and grade norms are given for 6,236 pupils in grades seven to twelve and in college.

EXAMINATION: FORM A—Price per package of 25 examination books, including Key and Age-Grade-Score Sheet and Percentile Graphs 80 cents net.

MANUAL OF DIRECTIONS—Price 15 cents net.

SPECIMEN SET—Price 25 cents postpaid.

WORLD BOOK COMPANY

YONKERS-ON-HUDSON, NEW YORK
2126 PRAIRIE AVENUE, CHICAGO

OTIS SELF-ADMINISTERING TESTS OF MENTAL ABILITY

DEvised BY ARTHUR S. OTIS
Author of Otis Group Intelligence Scale

A SERIES of tests designed for the range of grades from five to twelve, and for college freshmen. The examinations are distinguished by seven new features which make for economy of time and cost in testing, and enhance the value of the results.

1. *Self-administration.* The examiner has but to give a few initial directions, after which the students proceed with the test without interruption.
2. *Ease of scoring.* The answers to the questions appear in columns on the margins of three pages. With the key, the paper may be scored in about 45 seconds.
3. *Flexible time limit.* The examination can be given with the time limit of either 20 or 30 minutes depending upon accuracy desired and time available.
4. *Variety of test material.* The arrangement permits of the use of a large variety of questions which insure a more comprehensive measurement of mental ability.
5. *Ease of figuring IQ's.* By locating a point on a chart, an IQ can be found directly from the score and chronological age.
6. *Percentile graph.* A new form shows vividly the distribution of scores, and the Scale Chart provided with it simplifies the drawing of the curves.
7. *Chart to aid in classification.* An Interpretation Chart facilitates the division of a class or school into fast and slow moving groups taking account of brightness and chronological age as well as mental age.

INTERMEDIATE EXAMINATION: FORM A or FORM B. Price per package of 25 examination booklets, 1 Directions and Key, 1 Interpretation Chart and Percentile Graph, and 1 Class Record, 80 cents net

HIGHER EXAMINATION: FORM A or FORM B. Price per package of 25 examination booklets, 1 Directions and Key, 1 Interpretation Chart and Percentile Graph, and 1 Class Record, 80 cents net. SPECIMEN SET. Price 30 cents postpaid.

WORLD BOOK COMPANY

YONKERS-ON-HUDSON, NEW YORK
2126 PRAIRIE AVENUE, CHICAGO

Cross English Test

BY E. A. CROSS

*Professor of Literature and English
and Dean State Teachers College,
Greeley, Colorado*

THE purpose of this test is to obtain an accurate measure of the ability of high school or freshman college students to use correctly the common English forms. It does not test the student's knowledge of literature or the finer points of rhetoric. Its field is the sentence, the fundamental unit of composition, and it covers spelling, pronunciation, punctuation, grammatical forms, and sentence structure.

The test can be used as an aid to assigning marks and determining promotions, in diagnosing the special needs of students for certain kinds of instruction, in measuring the progress of an individual or a class, and for the classification of entering students into homogeneous groups; and it will serve as a college entrance examination.

The test is issued in three similar and equivalent forms. It has been made wholly objective by relying upon certain "key" errors which have been found to be evidence of general crudity or lack of knowledge in English. The test has been used in various preliminary forms for six years and tentative percentile norms for entering college freshmen are furnished.

Examination: Form A, Form B, or Form C. Price per package of 25 examinations, with 1 Manual of Directions, 1 Key, and 1 Class Record, \$1.20 net.

Specimen Set. Envelope containing 1 Examination and 1 Key of each Form, 1 Manual of Directions, and 1 Class Record. Price 25 cents postpaid.

WORLD BOOK COMPANY

YONKERS-ON-HUDSON, NEW YORK
2126 PRAIRIE AVENUE, CHICAGO

TWO IMPORTANT TESTS

By V. A. C. HENMON

Professor of Educational Psychology

Director of the School of Education

University of Wisconsin

LATIN

The result of seven years' experimentation. The words used are standard, being the common words from thirteen different beginners' books, Caesar, Cicero, and Vergil. The resulting tests form, therefore, the safest basis for grading pupils in any of the four years of Latin. A vocabulary examination and a sentence examination are contained in each test and both can be given in twenty minutes. The method of scoring is unique and simple. Each of the examinations is of equal difficulty, and offered to provide variety.

The tests, which are each two pages in extent, are put up in packages of 25 tests, including Directions for Administering and Scoring (with standard scores) and Class Record and Report to Author.

Test X is intended for research purposes, and use in school surveys.

Test 1.

Price per package 50 cents net.

Test 2.

Price per package 50 cents net.

Test 3.

Price per package 50 cents net.

Test 4.

Price per package 50 cents net.

Test X.

Price per package 50 cents net.

Specimen Set. An envelope containing 1 of each of the five tests, Directions, and Class Record. Price 10 cents postpaid.

FRENCH

Scientifically constructed standard tests, based on reliable norms, made after many years' investigation. A checking of the vocabularies of twelve recently and widely used first-year textbooks in French showed 448 words common to all the books. These words are taken as the standard vocabulary, and each word is assigned its proper weight. The three different tests are of the same grade of difficulty, so any one can be used first. The tests can be given at any time to students in any one of the four years of French, and will serve as the best method for grading attainment. Each test has a vocabulary and a sentence examination; both can be given in twenty minutes.

The tests, which are each two pages in extent, are put up in packages of 25 tests, including Directions for Administering and Scoring (with standard scores) and Class Record and Report to Author.

Test 1.

Price per package 50 cents net.

Test 2.

Price per package 50 cents net.

Test 3.

Price per package 50 cents net.

Test 4.

Price per package 50 cents net.

Specimen Set. An envelope containing 1 of each of the three tests, Directions, and Class Record. Price 10 cents postpaid.

WORLD BOOK COMPANY

YONKERS-ON-HUDSON, NEW YORK
2126 PRAIRIE AVENUE, CHICAGO

UNIVERSAL
LIBRARY



122 516

UNIVERSAL
LIBRARY